# Grundkurs Statistik für Historiker: T. I, Deskriptive Statistik

Thome, Helmut

Veröffentlichungsversion / Published Version
Themenheft / topical issue

**Empfohlene Zitierung / Suggested Citation:**
Thome, H. (1989). Grundkurs Statistik für Historiker: T. I, Deskriptive Statistik. *Historical Social Research, Supplement*, 2, 1-147. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-285945

HSR Supplement/Beiheft No. 2 (1989)
(Neudruck Dezember 1995)

# HELMUT THOME
# GRUNDKURS STATISTIK FÜR HISTORIKER
# TEIL I: DESKRIPTIVE STATISTIK

## Inhaltsverzeichnis

---

Editorial Staff: Rainer Metz (Assistant Editor), Petra Klein (Assistant), Claudia Strittmatter (Typesetter).

HSR was composed by the program *Satz* of *TUSTEP*. (*Tü*bingen *S*ystem of *Te*xtprocessing *P*rograms). HSR was printed by HUNDT-DRUCK (Cologne).

## Vorwort

Seit Mitte der siebziger Jahre hat die Historische Sozialforschung, die sich an den allgemeinen Methodenstandards einer empirisch-analytischen Wissenschaft orientiert, vor allem unter jungen Historikern, Doktoranden und Studenten zunehmend Interesse und Anerkennung gefunden. Das hat die geschichtswissenschaftlichen Ausbildungsgänge an den Hochschulinstituten der Bundesrepublik wenig beeinflußt. Lehrveranstaltungen über formale und quantifizierende Analysemethoden sind noch immer Ausnahmen.

Das Zentrum für Historische Sozialforschung (ZHSF) führt deshalb seit Jahren »Herbstseminare« durch, in denen neben anderen Inhalten zwei Grundkurse zur statistischen Methodenlehre angeboten werden. Das hier vorgelegte Skript zur deskriptiven Statistik entstand im Rahmen des Lehrprogramms zum »Grundkurs I«. Ein darauf aufbauendes Skript zum »Grundkurs II: Inferenzstatistik und Regressionsanalyse« wird voraussichtlich bis Ende 1989 druckreif vorliegen (siehe die hier im Anhang abgedruckte Themenübersicht).

Trotz des umfangreichen Angebots an »Einführungen in die Statistik für Sozialwissenschaftler« einen weiteren Einführungstext vorzulegen, scheint auf den ersten Blick überflüssig. Auf dem deutschen Buchmarkt ist jedoch nach unserer Kenntnis neben Armingers sehr guter, aber auch sehr gedrängter Darstellung (in Jarausch/Arminger/Thaller 1985) keine systematische Einführung in die statistische Methodenlehre erschienen, die speziell für Historiker verfaßt worden wäre. Die Erfahrung zeigt, daß sich Lernbarrieren im Zugang zu formalen Methoden am leichtesten überwinden lassen, wenn die entsprechenden Konzepte und Analyseverfahren anhand von Daten und Fragestellungen aus dem eigenen Fachgebiet vermittelt werden. Zwar sind einige Vielzweckbände über »Quantitative Methoden für Historiker« (oder ähnliche Titel) erschienen; in ihnen werden aber die eigentlich »statistischen« Methoden und Modellkonstruktionen viel zu knapp abgehandelt.

Statistik wird in diesem Skript als ein Instrumentarium präsentiert, mit dem Historiker (wie auch Wissenschaftler aus anderen Disziplinen) Daten analysieren, Informationen verdichten, Zusammenhänge und Strukturen in ihnen erkennen und, auf dieser Grundlage, theoretische Hypothesen explorieren oder testen können. Großer Nachdruck wird darauf gelegt, die Verknüpfung inhaltlicher (substanzwissenschaftlicher) Konzepte mit formal-statistischen Modellvorstellungen zu erörtern und die Anwendungsvoraussetzungen einzelner Verfahren zu klären.

Bei den Kursteilnehmern werden keine statistischen oder mathematischen Kenntnisse vorausgesetzt, die über das Niveau allgemeiner Schulbildung hinausgehen. Auswahl und Darstellung der verschiedenen Themen orientieren sich an folgenden Kriterien:

(1) Der angebotene »Stoff« sollte in einem Anfängerkurs innerhalb von 12 Doppelstunden zu bearbeiten sein. (Abschnitte, die bei einem ersten Lektüredurchgang ausgelassen werden können, sind mit einem Sternchensymbol (*) gekennzeichnet.) (2) Das Skript soll sich auch für das individuelle Studium außerhalb von Lehrveranstaltungen eignen. (3) Es soll den Zugang zur weiterführenden Literatur und zu komplexeren Verfahren (auf die fortlaufend hingewiesen wird) erleichtern (denn in der Forschungspraxis reichen die elementaren Verfahren, auch die in Teil II behandelten, häufig nicht aus).

Die Analysebeispiele sind mit dem Programmsystem SPSS$^{x}$ (Statistical Package for the Social Sciences), Version 3.1, ausgeführt worden. Die entsprechenden »Befehle« für die an Großrechnern installierte Fassung werden im Text zitiert. Das Skript ist jedoch bewußt auf die Erörterung statistischer Konzepte und Verfahren begrenzt; es bietet keine Einführung in die elektronische Datenverarbeitung.

Zur Durchführung der Analysen standen vier Datensätze zur Verfügung: (1) Die Abgeordneten der Frankfurter Nationalversammlung 1848/49 (von Heinrich Best). (2) Die Abgeordneten der Reichstage von 1867 - 1918 (Heinrich Best). (3) Biographische Daten der SPD-Reichstagskandidaten 1898 - 1912 (Wilhelm H. Schröder). (4) Wahlkreisdaten zu den Reichstagswahlen von 1898 - 1912 (Wilhelm H. Schröder).

Ich danke den Mitarbeitern des ZHSF sowie Heinrich Best, Jörg Blasius, Steffen Kühnel, Herbert Odenthal und Kurt Sombert, die Teile des Skripts gelesen und mit hilfreichen Kommentaren versehen haben. Ralph Ponemereo und Kurt Sombert danke ich außerdem für ihre großzügige Unterstützung bei den EDV-Arbeiten.

Köln, Mai 1989 — Helmut Thome

## Anmerkung zum Neudruck

Da die erste Auflage rasch vergriffen war, wurde ein Neudruck erforderlich. Sein Text weicht lediglich durch einige kleinere Fehlerkorrekturen von dem des ersten Drucks ab.

Köln, Dezember 1995 — H. T.

## Kapitel 1

## Merkmalsdimensionen und Meßniveaus

Systematisches Forschen beginnt mit einer theoretisch angeleiteten Problemdefinition. Der Forscher muß sich über seine Fragen klar sein, er muß seine Untersuchungsobjekte (Individuen oder Kollektive) bestimmen und die Merkmalsdimensionen festlegen, die er erheben will. Erst dann kann er »Daten« sammeln, kann er »messen«, also die Ausprägungen (»Werte«) ermitteln, die seine Untersuchungsobjekte (Untersuchungseinheiten) auf den ausgewählten Merkmalsdimensionen aufweisen. Sowohl Individuen als auch Kollektive können als Untersuchungseinheiten dienen. In den verschiedenen Datensätzen für diesen zweiteiligen Kurs sind die Abgeordneten der Frankfurter Nationalversammlung (1848/49), die Reichstagskandidaten der SPD für die Wahlen von 1898 bis 1912, die Reichstagsabgeordneten aller Parteien von 1867 bis 1918 sowie die Wahlkreise zu den Reichstagswahlen unsere Untersuchungseinheiten; »Religionsbekenntnis«, »Erlernter Beruf« der Reichstagskandidaten, »Grad der Industrialisierung«, »Bevölkerungsanteil in Gemeinden unter 2000 Einwohnern« der Wahlkreise sind einige der erhobenen Merkmalsdimensionen.

»Messen« bedeutet formal das Zuordnen von Zahlen (oder sonstigen Symbolen) zu den empirischen Merkmalsträgern nach bestimmten Regeln. Diese Regeln sollen sicherstellen, daß die Objekte durch die Zahlen »strukturtreu« abgebildet werden (wird gleich erläutert). Durch den Meßvorgang werden Merkmalsdimensionen zu »Variablen«, die einer quantitativen Analyse zugänglich sind. Die jeweilige Zuordnung von Zahlen zu bestimmten Merkmalsausprägungen wird in einem »Codeplan« festgehalten. *Abb. 1.1* bringt einen Auszug aus dem Codeplan für den Datensatz der SPD-Reichstagskandidaten. In diesem Falle ist der Codeplan in eine umfassendere Datendokumentation eingearbeitet, die auch Erläuterungen zu den einzelnen Variablen enthält.

Das Messen kann auf unterschiedlichen »Niveaus« (Skalenniveaus) erfolgen (Skala = Achse mit zugeordneten numerischen Ausprägungen). Wo Zahlen vorliegen, kann man Relationen herstellen, z. B. gleich/ungleich oder größer/kleiner. Die Meßniveaus werden danach unterschieden, welche Arten von Relationen zwischen den Zahlenwerten, die man den Ausprägungen einer Merkmalsdimension zugeordnet hat, als »zulässig« deklariert werden. Die Zulässigkeit richtet sich danach, welche Relationen zwischen den Untersuchungsobjekten als inhaltlich sinnvoll anzusehen sind. Zum Beispiel ist es für den Sozialwissenschaftler in der Regel nicht sinnvoll, zwischen den Kategorien (Ausprägungen) der Merkmalsdimension »Konfessionszugehörigkeit« Relationen vom Typ »größer/kleiner« (in

Abb. 1.1: Auszug aus Datendokumentation/Codebuch zum Handbuch Schröder: Sozialdemokratische Reichtagsabgeordnete und Reichstagskandidaten 1898 - 1918



C: 0 = nein
1 = ja, in Emigration

**FBRDDR**
**B: Relevante Funktion in Politik und Verwaltung nach 1945 in den Besatzungszonen bzw. in der BDR und der DDR (Erhebung 1987)**
E: Nur überregionale und regionale Funktionen wurden erfaßt.
M: nominal, F1.0
C: 0 = nein
1 = ja

**MDR**
**B: Mitglied des Reichstages (Erhebung 1974)**
E: Die Frage ist unabhängig von der konkreten Dauer der Mitgliedschaft. Die Mitglieder der verfassungsgebenden Nationalversammlung 1919/20 gelten als gleichwertig.
M: nominal, F1.0
C: 0. NIE MDR
1. NUR VOR 1918 MDR
2. NUR NACH1918 MDR
3. VOR UND NACH 1918 MDR

**MDR4**
**B: Mitgliedschaft im Deutschen Reichstag 1867-1933 (Erhebung 1987)**
E: MDR schließt die Mitgliedschaft im Norddeutschen Reichstag und in der Deutschen Nationalversammlung ein.
M: nominal, F1.0
C: 0 = unzutreffend
1 = nur im Kaiserreich
2 = nur in der Weimarer Republik
3 = im Kaiserreich und in der Weimarer Republik

**MDL4**
**B: Mitgliedschaft in einem deutschen Landtag 1871-1933 (Erhebung 1987)**
E: MDL schließt alle 28 Länderparlamente ("Landtag", "Abgeordnetenhaus", "Bürgerschaft", "Volkstag", "Zweite Kammer", "Landesversammlung" etc., einschließlich der verfassungsgebenden Landesversammlungen 1919) im Deutschen Reich ein, die u.a. aus "allgemeinen" Wahlen hervorgegangen sind. Aufgrund der besonderen politischen Situation in der Freien Stadt Danzig bestand der "Volkstag" noch über 1933 hinaus; im Falle von Danzig schließt daher ausnahmeweise MDL die Mitgliedschaft im "Volkstag" bis zu seiner Auflösung im Jahre 1938 ein.
M: nominal, F1.0
C: 0 = unzutreffend
1 = nur im Kaiserreich
2 = nur in der Weimarer Republik
3 = im Kaiserreich und in der Weimarer Republik

**RTKDT**
**B: Kandidatur zum Deutschen Reichstag (Erhebung 1987)**
E: Bis 1918 wurden alle nachweisbaren und "offiziellen" sozialdemokratischen Kandidaten bei den stattgefundenen Haupt-, Stich-, Ersatz- und Nachwahlen im Rahmen des Mehrheitswahlrechtes berücksichtigt. 1919-1933 wurden alle Kandidaten berücksichtigt, die im Rahmen des Verhältniswahlrechtes auf den Kandidatenlisten der Wahlkreise für die (M)SPD bzw. USPD nominiert worden waren.
M: nominal, F1.0
C: 0 = unzutreffend
1 = nur im Kaiserreich
2 = nur in der Weimarer Republik
3 = im Kaiserreich und in der Weimarer Republik

Symbolen: > oder <) herzustellen. Lediglich die Relationen »gleich/ungleich« (= , ≠) sind hier vertretbar. Allerdings könnten andere Merkmalsdimensionen, die mit der Konfession eng zusammenhängen (z. B. Umfang der vorgeschriebenen Kulthandlungen), sehr wohl auch weitere Relationen (im Beispiel: > , < ) zulassen. Es hängt also von der genauen Definition der Merkmalsdimension (Variablen) und der in sie eingegangenen theoretischen Perspektive ab, welche Relationen (und damit welches Meßniveau) als sinnvoll anzusehen sind.

Im Hinblick auf die Untersuchungsobjekte und deren Relationen auf einer Merkmalsdimension spricht man von einem »empirischen Relativ«. Das »numerische Relativ« der zugeordneten Zahlen soll das jeweilige empirische Relativ strukturgleich abbilden. Das heißt, die Beziehungen der Objekte zueinander (z. B. »größer/kleiner«) müssen durch die Beziehungen der jeweils zugeordneten Zahlen korrekt wiedergegeben werden. Dazu müssen die Kriterien der Eindeutigkeit, Ausschließlichkeit und Vollständigkeit (siehe Schröder 1987, Kap. IV-1) erfüllt sein.

Im allgemeinen unterscheidet man vier Meß- oder Skalenniveaus: Nominal-, Ordinal-, Intervall- und Ratioskala. Von den Skalenniveaus hängt ab:

(a) welche Transformationen X' = T(X) zulässig sind, um eine bereits vorliegende Skala (X) zu einer neuen, äquivalenten Skala (X') zu transformieren
(b) welche numerischen Operationen mit den Werten $x_1$, $x_2$,...,$x_m$ einer Skala X zulässig sind, um **Vergleiche** zwischen den Untersuchungsobjekten auf einer bestimmten Merkmalsdimension vorzunehmen.

Der Typ algebraischer Operationen (wie z. B. Addieren, Multiplizieren, Logarithmieren), der legitimerweise bei Skalentransformationen anwendbar ist, bei denen alle Werte einer Skala **verändert** werden, ist von den zulässigen Rechenoperationen zu unterscheiden, mit denen bei gegebenem Skalenniveau einzelne Untersuchungsobjekte hinsichtlich ihrer Skalenwerte **verglichen** werden.

Viele statistische Analyseverfahren sind speziell für ein ganz bestimmtes Skalenniveau definiert und für andere überhaupt nicht zulässig. Wir wollen nun für die einzelnen Skalenniveaus Beispiele nennen und die jeweils zulässigen Skalen-Transformationen und vergleichenden Rechenoperationen erläutern.

**Nominalskala:**

Hier haben wir schon das Religionsbekenntnis der Reichstagskandidaten als Beispiel erwähnt. In unserem Datensatz sind dafür folgende Ausprägungen (Kategorien) definiert: keine Angaben(0); evangelisch (1); evangelisch, später dissident (2); katholisch (3); katholisch, später dissident (4);

jüdisch (5); jüdisch, später dissident (6); dissident, frühere Konfession unbekannt (7); sonstiges (8). Die in Klammern gesetzten Zahlen sind ihnen laut Codeplan zugeordnet. Reichstagskandidaten, die die gleiche Konfession haben, erhalten auf dieser Variablen den gleichen Wert, Kandidaten mit unterschiedlicher Konfession ungleiche Werte. Es sind nur die Relationen gleich/ungleich sinnvoll. Wir sind zwar gewohnt, mit Zahlen Rechenoperationen wie Addieren oder Subtrahieren auszuführen; in unserem Beispiel fungieren die Zahlen aber lediglich als Kürzel für bestimmte Kategorien. Für jeden Skalenwert (für jede Kategorie) kann man auszählen, wie häufig er in der Menge der Untersuchungsobjekte vorkommt (»Häufigkeitsauszählung«). Aber Rechenoperationen innerhalb der Skala (z. B. Subtrahieren eines Wertes von einem anderen), können nicht inhaltlich interpretiert werden.

Statt der Zahlen hätte man auch andere Kürzel verwenden können (z. B. einzelne Buchstaben), doch bei der EDV versucht man möglichst viele Inhalte durch Zahlen bzw. Zahlenkombinationen auszudrücken. Man muß sich eben »merken«, d. h., man muß im Codeplan dokumentieren, »wofür« die Zahlen stehen, und welche Relationen zwischen ihnen inhaltlich sinnvoll sind. In dem obigen Beispiel hätten auch andere Zahlen verwendet werden können. Das heißt, die dort vorgelegte Skala könnte »transformiert« werden. Zugelassen wären alle Transformationen, die umkehrbar eindeutig sind, so daß ungleiche Zahlen nicht in gleiche Zahlen überführt und gleiche Zahlen nicht in ungleiche verwandelt werden. Statt der »1« für »evangelisch« hätte man beispielsweise auch die »5« einsetzen können, statt der »3« für »katholisch« die »4«, statt der »7« die »16« usw. Zu fordern ist lediglich, daß auch bei Anwendung der neuen Skala Kandidaten mit ungleichen Variablenwerten verschiedenen Konfessionen angehören und Kandidaten mit gleichen Variablenwerten die gleiche Konfession haben.

**Ordinalskala** (Rangskala):

Ein Beispiel hierfür ist das Niveau der Schulbildung der Reichstagskandidaten mit den Kategorien »niedrig« (1) »mittel« (2) »hoch« (3). Neben den Relationen gleich/ungleich sind jetzt auch die Relationen größer/kleiner sinnvoll definierbar. Die Rangrelationen der Skala müssen den Rangrelationen der Merkmalsausprägungen entsprechen. Aber die Größe der numerischen Abstände zwischen den Rangplätzen ist nicht inhaltlich interpretierbar. Die substantielle Differenz zwischen den Rangplätzen 2 und 3 kann genauso groß, kleiner oder größer sein als die Differenz zwischen den Rangplätzen 1 und 2. Das mag sowohl in der »Natur« der Merkmalsdimension als auch in der Unvollkommenheit unserer Meßinstrumente begründet sein. Subtrahieren, Dividieren oder Multiplizieren einzelner Rangwerte miteinander sind folglich nicht geeignet, Größenverhältnisse zwischen den Untersuchungsobjekten auszudrücken.

Häufigkeitsauszählungen sind natürlich möglich.

Zulässig sind alle Skalentransformationen, die die Rangordnung der Elemente unverändert lassen (»monotone« Transformationen), z. B. das Quadrieren: $X' = X^2$. Diese monotonen Transformationen enthalten als Untermenge alle Transformationen, die auf den höheren Skalenniveaus zulässig sind (siehe unten).

**Intervallskala:**

Ein Beispiel hierfür ist das Geburtsjahr. Die Relationen $(=, \neq)$ und $(>, <)$ sind nicht nur für den Vergleich **einzelner** Skalenwerte zugelassen, sondern auch auf den Vergleich von Differenzen, also auf ein Zahlenquadrupel $(x_1\text{-}x_2)$ und $(x_3\text{-}x_4)$ anwendbar. Das heißt, Subtrahieren (und Addieren) der Variablenwerte miteinander sind nun sinnvoll, da die Abstände zwischen den Merkmalsausprägungen quantitativ eindeutig bestimmt sind; gleiche Zahlen**differenzen** bedeuten inhaltlich (hinsichtlich der jeweiligen Merkmalsdimension) gleich große Abständen zwischen den Objekten, denen die Zahlen zugeordnet sind.

Zulässig sind **lineare** Skalentransformationen: $X' = aX + b$, $a > 0$.

Ein berühmtes Beispiel für die Transformation von Intervallskalen sind die Temperaturskalen. So wird die »Celsius«-Skala, C°, mit $a = 1{,}8$ und $b = 32$ in die »Fahrenheit«-Skala, F°, überführt:

| C° | F° | $Z^\circ = (C^\circ)^2$ |
|---|---|---|
| 0 | 32 | 0 |
| 10 | 50 | 100 |
| 30 | 86 | 900 |
| 100 | 212 | 10000 |

Das Verhältnis irgendwelcher **Differenzen** zwischen beliebigen Temperaturwerten der einen Skala ist gleich dem Verhältnis der entsprechenden Differenzen der anderen (äquivalenten) Skala. Auf der Celsiusskala ist z. B. (100 - 30)/(30 - 10) = 3,5. Für die Fahrenheitskala erhalten wir entsprechend: (212 - 86)/(86-50) = 3,5. Hätten wir eine nicht-lineare, z. B. eine quadratische Transformation $Z^0 = (C^\circ)^2$ vorgenommen, wäre eine nicht-äquivalente Skala entstanden: (10000-900)/(900-100) ≠ 3,5.

Es wäre übrigens nicht sinnvoll zu sagen, es sei bei einer Temperatur von 20 Grad Celsius »doppelt so warm« wie bei 10 Grad Celsius. Die Sinnlosigkeit dieser Operation zeigt sich, wenn wir diese Celsius-Werte in Fahrenheit-Werte transformieren und ins Verhältnis zueinander setzen: (68° F/50° F) ≠ 2. Es hat also keinen Sinn, bei einer Skala ohne »natürlichen« Nullpunkt Verhältnisse zweier **einzelner** Skalenwerte $x_1/x_2$ zu bilden. Das bringt uns zum letzten Skalentyp, der

**Ratio-** oder **Verhältnisskala,**

die zusätzlich zu den Eigenschaften der Intervallskala auch noch über einen »natürlichen« Nullpunkt verfügt, der theoretisch interpretiert werden kann. Beispiele sind das Monatseinkommen oder das Lebensalter. Es sind alle Relationen zulässig, die auch für die Intervallskala sinnvoll sind, plus der Bildung von Verhältnisgrößen zwischen beliebigen Skalenwerten (nicht nur zwischen Differenzen von Skalenwerten). So kann man z. B. feststellen, daß jemand, der 4000 DM verdient, doppelt soviel verdient wie jemand, der 2000 DM erhält; ein 60jähriger ist doppelt so alt wie ein 30jähriger. Diese Beispiele zeigen aber auch, wie umstritten die Zuerkennung dieses Skalenniveaus (oder auch nur des Intervallskalenniveaus) sein kann. In welchem Sinne ist ein 60jähriger doppelt so alt wie ein 30jähriger - biologisch? (kaum), soziologisch? (fragwürdig)[1].

Auch die Einkommensdifferenz zwischen 4000 und 4200 DM muß nicht (kann aber) das gleiche bedeuten wie die Differenz zwischen 2000 und 2200 DM. Theoretische Erwägungen müssen hier das Skalenniveau festlegen. Manche Forscher (und Tarifpolitiker) möchten z. B. Einkommensdifferenzen nur dann als gleich(wertig) betrachten, wenn sie gleiche prozentuale Veränderungen ausdrücken. In diesem Sinne wäre zum Beispiel ein Anstieg von 4000 auf 4400 DM »gleich« dem Anstieg von 2000 auf 2200 DM; der Zuwachs wäre in beiden Fällen 10 Prozent.

(*) Zuwachsraten kann man durch Logarithmieren der Ausgangsskala abbilden. So ist z. B. log(4000)= 3,602; log(4400)= 3,643; log(4400) - log(4000) = 0,041. Die gleiche Differenz erhält man durch log(2200) - log(2000) = 3,342 - 3,301 = 0,041. Gleiche Log-Differenzen bedeuten somit gleiche Einkommensproportionen. Zu beachten ist, daß der Logarithmus keine lineare Transformation der Ausgangsdaten darstellt. Wenn sich der Log-Wert um den Faktor »a« vervielfacht, vervielfachen sich die Einkommensbeträge um die Potenz »a«:

(1-1) $\log x_2 = a \cdot \log x_1$

$10^{\log x_2} = 10^{a \cdot \log x_1}$

$x_2 = \left(10^{\log x_1}\right)^a$

$x_2 = x_1^a$

[1] Hier wird ein weiterreichendes wissenschaftslogisches Problem berührt, das wir in diesem Skript nicht behandeln können, nämlich die Differenz von »latenten« und »manifesten« Variablen, von theoretischem Konstrukt und empirischem Indikator. (Siehe hierzu Einführungen in die Wissenschaftslogik, z. B. Giesen/Schmid 1976).

Ebenso gilt: Wenn sich die Differenz der Log-Werte um den Faktor »c« vervielfacht, vervielfacht sich der Quotient der Einkommensbeträge um die Potenz »c«.

Wenn eine Ratioskala vorliegt, kann sie nur durch eine proportionale Transformation $X' = aX$, $a > 0$, in eine neue, äquivalente Ratioskala transformiert werden. Dadurch verändern sich zwar die Abstände zwischen den einzelnen Werten, aber die Verhältnisse entsprechender Differenzen bleiben gleich.

Je höher das Skalenniveau, um so geringer also die Menge der erlaubten Transformationen, aber um so »höher« der Typ zugelassener Rechenoperationen, mit denen Relationen zwischen den empirischen Objekten sinnvoll charakterisiert werden können.

Statistische Analyseverfahren implizieren häufig Skalentransformationen und relationieren Elemente der Objektmenge mittels Rechenoperationen. Deshalb ist jeweils zu prüfen, welches Meßniveau dabei vorausgesetzt wird. Skalen höheren Niveaus lassen sich stets wie Skalen niedrigeren Niveaus behandeln, das Umgekehrte gilt nicht.

Intervallskala und Ratio-Skala faßt man unter der Bezeichnung **metrische Skalen** zusammen; nominale und ordinale Variablen nennt man auch **topologische** Variablen. Bei Nominalskalen spricht man gelegentlich von »qualitativen« oder »kategorialen« Variablen.

Eine Sonderrolle spielen

**dichotome Variablen,**

also Variablen, die nur zwei Ausprägungen aufweisen, wie z. B. das Geschlecht. Inhaltlich liegt hier sicherlich eine qualitative Variable vor. Wenn man »weiblich« mit 1 und »männlich« mit 0 kodiert, ist dennoch das arithmetische Mittel sinnvoll interpretierbar, nämlich als Anteilswert. Wenn von 100 Personen 55 den Wert »1« aufweisen (also Frauen sind) und 45 den Wert »0« haben, so ergibt sich daraus eine Summe von 55 und ein arithmetisches Mittel von 0,55. Man kann sogar die Meinung vertreten, daß es sich bei dichotomen (binär kodierten) Variablen um Ratioskalen handele, da die eine der beiden Ausprägungen als natürlicher Nullpunkt im Verhältnis zur anderen interpretiert werden kann: den Männern z. B. fehlt das Merkmal »weiblich«, sie haben es mit der Ausprägung »Null« (unbeschadet biologischer und psychologischer Theorien, die da weiter differenzieren mögen). Dichotome Variablen können in der statistischen Analyse häufig wie metrische Skalen behandelt werden, aber das ist gelegentlich umstritten. Auch gibt es einige Verfahren und Kennwerte speziell für dichotome Variablen.

Auch Variablen, die ursprünglich mehr als zwei Ausprägungen aufweisen, können im Verlauf der Analyse zu dichotomen Variablen zusammengefaßt werden, so z. B. die Variable »Einkommen«, indem man nur noch zwischen Beziehern von niedrigem (unterdurchschnittlichem) und hohem Einkommen unterscheidet.

# Kapitel 2

# Darstellung univariater Häufigkeitsverteilungen

## 2.1 Datenmatrix

Bevor man die Daten statistisch analysieren kann, müssen sie in geeigneter Weise geordnet werden: als Zahlen in einem rechteckigen Schema (»Matrix«), das aus n Zeilen und m Spalten besteht. Die n Zeilen repräsentieren die n Untersuchungseinheiten (Objekte, Merkmalsträger), die m Spalten die erhobenen Merkmale (Variablen), deren Zahl in der Regel kleiner als m ist. Das Schema der **Datenmatrix** läßt sich also wie folgt skizzieren:

**Abb.2.1**: Schema einer Datenmatrix

| | Variablen | | | | | |
|---|---|---|---|---|---|---|
| | $V_1$ | $V_2$ | ... | $V_j$ | ... | $V_m$ |
| $UE_1$ | $W_{11}$ | $W_{12}$ | ... | $W_{1j}$ | ... | $W_{1m}$ |
| $UE_2$ | $W_{21}$ | $W_{22}$ | ... | $W_{2j}$ | ... | $W_{2m}$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ |
| $UE_i$ | $W_{i1}$ | $W_{i2}$ | ... | $W_{ij}$ | ... | $W_{im}$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ |
| $UE_n$ | $W_{n1}$ | $W_{n2}$ | ... | $W_{nj}$ | ... | $W_{nm}$ |

Die Symbole $W_{ij}$ repräsentieren den Wert, der für eine bestimmte Untersuchungseinheit i (i = 1, 2,...,n) bei einer bestimmten Variablen j (j = 1, 2,...,k) beobachtet bzw. registriert (»gemessen«) wurde. In der Regel ist die Zahl der Variablen (»k«) kleiner als die Zahl der Spalten (»m«), da einige Variablen, wie z. B. Lebensalter, zwei-oder mehrstellige Werte (Ausprägungen) aufweisen. Bei der statistischen Auswertung der Datenmatrix muß also stets angegeben werden, welche Spalten eventuell für eine Variable zusammenzufassen sind. (Alternativ hierzu könnte man die Spalten von vornherein als mehrstellig definieren und ihre Zahl gleich der Zahl der Variablen setzen.) Gelegentlich richtet man auch Leerspalten ein, um die Variablen übersichtlicher zu gruppieren.

Um das Schema der *Abb. 2.1* zu konkretisieren bringen wir in *Abb. 2.2* einen Auszug aus der Datenmatrix der SPD-Reichstagskandidaten.

Eine solche Datenmatrix kann man über Lochkarten (für jede Untersuchungseinheit eine oder mehrere) oder über den Bildschirm (Zeile für Zeile) in den Computer eingeben. Auch Buchstabenkombinationen, z. B. Namen von Untersuchungseinheiten, können in die Datenmatrix aufgenommen (und bei numerischen Analysen ohne weiteres überlesen) werden. In unserem Beispiel tauchen bestimmte Politiker-Namen mehrmals auf, wenn sie in mehreren Wahlkreisen und/oder bei mehreren Wahlen kandidierten. (In diesem Datensatz sind nicht die Kandidaten, sondern die Kadidaturen die Untersuchungseinheiten.) »Alphanumerische« Zeichen, also Zeichenkombinationen, die (auch) Buchstaben enthalten, können in SPSS$^x$ automatisch kodiert werden. (Man spricht in diesem Zusammenhang von »String-Variablen«.) Das ist vor allem nützlich bei der Erhebung von Merkmalsdimensionen wie z.B. »Beruf«, deren Ausprägungen man nicht im vorhinein kategorisieren möchte.

In den ersten Spalten einer Datenmatrix sollte stets Platz gelassen werden für eine oder mehrere Zahlenkombinationen, mit denen die jeweilige Untersuchungseinheit identifiziert werden kann. In der Regel empfiehlt es sich auch, das Datum der Erhebung, besondere Umstände bei der Datenerfassung und Quellencharakteristiken in die Datenmatrix aufzunehmen.

## 2.2 Häufigkeitsverteilungen

Einer der ersten Schritte in der Datenanalyse ist das Auszählen von Häufigkeiten: Man möchte wissen, wie häufig die einzelnen Werte der Variablen in der Menge der Untersuchungsobjekte vorkommen. Indem man die Häufigkeiten für alle Werte einer Variablen zusammenstellt, erhält man eine **Häufigkeitsverteilung**. Sie ist also die Zuordnung von Merkmalsausprägungen (Variablenwerten) zu der beobachteten Häufigkeit ihres Vorkommens in einer Menge von Untersuchungseinheiten. Bei Nominal- und Ordinalskalen bereitet diese Zusammenstellung keinerlei Schwierigkeiten, da die Skalen in der Regel nur eine geringe Zahl »diskreter«, auf dem Zahlenstrahl deutlich voneinander getrennter Werte aufweisen. Man bezeichnet sie deshalb auch als »diskrete« Variablen. Metrische Variablen (wie z. B. Einkommen oder Alter) sind »eigentlich« ebenfalls diskrete Größen, da die Messung nicht so fein ist, daß jeder beliebige Punkt auf dem Zahlenstrahl besetzt werden könnte. Dennoch behandelt man sie häufig als »kontinuierliche« (oder »stetige«) Variablen. Um die Häufigkeitsverteilung auch in diesen Fällen übersichtlich zu gestalten, faßt man benachbarte Ausprägungen zu »Klassen« oder »Gruppen« mit bestimmter Intervallbreite zusammen. Folglich spricht man auch von »klassierten« oder »gruppierten« Daten.

Abb. 2.2: Auszug aus der Datenmatrix der Reichstagskandidaten der SPD

| SYSTEM | SER | CAND | BIOND1 | JAHR | NUMMER | ALPHA | STAMMNR | NAME | SEX | GEBJAHR | TODJAHR | TOD | HERKUNFT | BERUFWAGE | REIFGRAD | REIFGRAD | MILITAER | BILDUNG | ERLBERUF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 4 | 1898 | 48 | 2 | 178 | BRAUN OTTO | 1 | 872 | 955 | 1 | 6 | 3 | 3 | 0 | 1 | | 9 |
| 1 | 1 | 7 | 4 | 1898 | 131 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 1 | 4 | 1898 | 125 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 1 | 4 | 1898 | 309 | 19 | 118 | SCHNELL FRANZ | 1 | 860 | 923 | 1 | 166 | 0 | 2 | 0 | 1 | | 40 |
| 1 | 1 | 2 | 4 | 1898 | 126 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 3 | 4 | 1898 | 127 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 4 | 4 | 1898 | 128 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 5 | 4 | 1898 | 129 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 6 | 4 | 1898 | 130 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 1 | 1 | 1898 | 382 | 2 | 172 | BRAUN AUGUST | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 542 |
| 1 | 1 | 1 | 3 | 1898 | 154 | 8 | 152 | HOFER ADOLF | 1 | 868 | 935 | 1 | 15 | 4 | 0 | 0 | 5 | | 202 |
| 1 | 1 | 2 | 3 | 1898 | 155 | 8 | 152 | HOFER ADOLF | 1 | 868 | 935 | 1 | 15 | 4 | 0 | 0 | 5 | | 202 |
| 1 | 1 | 8 | 4 | 1898 | 132 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 9 | 4 | 1898 | 133 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 9 | 4 | 1898 | 134 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 1 | 3 | 1898 | 72 | 5 | 7 | EBHARDT ERNST | 1 | 849 | 0 | 0 | 20 | 4 | 0 | 0 | 6 | | 542 |
| 1 | 1 | 9 | 4 | 1898 | 135 | 8 | 6 | HAASE HUGO | 1 | 863 | 919 | 4 | 12 | 3 | 4 | 0 | 6 | | 201 |
| 1 | 1 | 1 | 3 | 1898 | 334 | 20 | 59 | STORCH FRANZ | 1 | 863 | 0 | 0 | 73 | 0 | 0 | 0 | 0 | | 13 |
| 1 | 1 | 2 | 3 | 1898 | 335 | 20 | 59 | STORCH FRANZ | 1 | 863 | 0 | 0 | 73 | 0 | 0 | 0 | 0 | | 13 |
| 1 | 1 | 3 | 3 | 1898 | 336 | 20 | 59 | STORCH FRANZ | 1 | 863 | 0 | 0 | 73 | 0 | 0 | 0 | 0 | | 13 |

In *Abb. 2.3 a-e* präsentieren wir für jedes Meßniveau eine Häufigkeitsverteilung aus dem Datensatz der SPD-Kandidaten für die Wahl zum Reichstag in 1912. In SPSS$^x$ enthält man den entsprechenden Ausdruck mit dem Befehl

```
FREQUENCIES
  VARIABLES = (hier folgt eine Liste der Variablen, deren Häu-
              figkeitsverteilung man sehen möchte)
  /BARCHART = FREQU
  /STATISTICS = MODE MEDIAN MEAN RANGE STDDEV
              VARIANCE
```

Das Subkommando BARCHART erzeugt Säulendiagramme, die wir in Abschn. 2.3 besprechen; über das Subkommando STATISTICS erhält man statistische Kennzahlen, die in Kap. 3 erörtert werden. Es können noch weitere Subkommandos in den FREQUENCIES-Befehl eingebaut werden, die uns an dieser Stelle aber nicht interessieren sollen.

Vor den Prozedurbefehl FREQUENCIES wurde noch das Kommando

```
MISSING VALUES KONFESSION (0)
```

eingesetzt, mit dem diejenigen Werte als »fehlend« (oder invalide) gekennzeichnet werden, die bei der Berechnung von Statistiken nicht berücksichtigt werden sollen. Falls man sie bei einer späteren Prozedur doch berücksichtigen möchte, kann das - bei der Pozedur FREQUENCIES - mit dem zusätzlichen Subkommando /MISSING = INCLUDE, bei anderen Prozeduren mit einem OPTIONS-Kommando bewerkstelligt werden. Wird /MISSING = INCLUDE bei FREQUENCIES nicht angegeben, sorgt in diesem Falle die Voreinstellung des Systems dafür, daß die zuvor als »Missing« deklarierten Werte dennoch in den Häufigkeitstabellen erscheinen, aber beim Berechnen der Statistiken ausgelassen werden.

Bei fast allen Datenerhebungen kommt es vor, daß zumindest bei einigen Merkmalsdimensionen für einige Fälle keine »Beobachtungen« vorliegen. Solange diese Fälle auf anderen Variablen valide Werte aufweisen, wäre es unsinnig, sie aus der Datenmatrix auszuschließen. Statt dessen sollte den fehlenden Beobachtungen im Datensatz eine besondere Codeziffer zugeordnet werden. In unserem ersten Beispiel (siehe *Abb. 2.3 a*) fehlen von 80 Reichstagskandidaten Angaben zum Religionsbekenntnis. Für sie ist schon bei der Konstruktion des Datenfiles die besondere Kategorie (»Keine Angabe«) mit dem Wert »0« eingerichtet worden. Es kann natürlich auch ein anderer Wert hierfür angegeben werden; er darf aber nicht mit einem Wert für gültige Beobachtungen identisch sein. Fehlende Werte sollten zumindest bei der Konstruktion des Datensatzes und bei einer ersten Häufigkeitsauszählung von den Kategorien »weiß nicht«,

»Klassifikation nicht eindeutig« und ähnlichem getrennt bleiben. Bei späteren Auswertungen kann man jeweils entscheiden, ob oder wie man diese Fälle berücksichtigen will. Ein höheres als das nominale Meßniveau ist bei einer statistischen Analyse in der Regel nur erreichbar, wenn Kategorien wie »unbekannt«, »andere«, »nicht zutreffend« ausgeschlossen werden. (Zum Problem fehlender Werte in der Datenanalyse siehe Kap. 13 in Teil II dieses Grundkurses.)

*Abb. 2.3 a* zeigt die Verteilung des Religionsbekenntnisses der Reichstagskandidaten. In der ersten Spalte (»Value«) stehen die Codeziffern für die einzelnen Ausprägungen, denen, wie oben erläutert, bei Nominalvariablen keine numerische Bedeutung zukommt. Die zweite Spalte (»Frequencies«) enthält die **absoluten Häufigkeiten**: die Menge der Fälle, denen ein bestimmter »Wert« zugewiesen wurde.

Bezeichnet man mit $f(x_i)$, $i = 1,...,m$, die absolute Häufigkeit der Ausprägung i einer Variablen X, die m Ausprägungen oder Klassen aufweist, so gilt:

$$(2\text{-}1) \qquad f(x_1) + f(x_2) + .... + f(x_m) = \sum_{i=1}^{m} f(x_i) = n$$

wobei n für die Gesamtzahl der Untersuchungseinheiten steht.

Die **relative Häufigkeit**, $f_r(x_i)$, ist definiert durch:

$$(2\text{-}2) \qquad f_r(x_i) = \frac{f(x_i)}{n}$$

Die Summe der relativen Häufigkeiten muß offensichtlich 1 ergeben:

$$(2\text{-}3) \qquad \sum_{i=1}^{m} f_r(x_i) = 1$$

Prozentanteile erhält man, indem man die relativen Häufigkeiten mit 100 multipliziert. Sie stehen in der dritten und vierten Spalte der Häufigkeitsverteilung (s. *Abb. 2.3 a*). Spalte 3 (»Percent«) enthält die Prozentuierungen auf der Basis aller Fälle, einschließlich der Fälle mit fehlenden Beobachtungen (»Missing«). In der vierten Spalte stehen die Prozentangaben, die sich ergeben, wenn man die fehlenden Beobachtungen nicht mitzählt. Die letzte Spalte (»Cum Percent«), enthält die sog. **kumulierten Häufigkeiten**. Sie sind nur bei ordinalem oder »höherem« Meßniveau aussagekräftig, so daß wir dieses Konzept weiter unten erläutern werden.

Abb. 2.3: Univariate Häufigkeitsverteilungen

a) Nominalvariable: Religionsbekenntnis (Reichstagskandidaten 1912

| VALUE LABEL | VALUE | FREQUENCY | PERCENT | VALID PERCENT | CUM PERCENT |
|---|---|---|---|---|---|
| EVANGELISCH | 1 | 83 | 20.9 | 26.2 | 26.2 |
| EV-->DISSIDENT | 2 | 84 | 21.2 | 26.5 | 52.7 |
| KATHOLISCH | 3 | 34 | 8.6 | 10.7 | 63.4 |
| KATH-->DISSIDENT | 4 | 30 | 7.6 | 9.5 | 72.9 |
| JUEDISCH | 5 | 11 | 2.8 | 3.5 | 76.3 |
| JUED-->DISSIDENT | 6 | 8 | 2.0 | 2.5 | 78.9 |
| DISSIDENT | 7 | 65 | 16.4 | 20.5 | 99.4 |
| SONSTIGES | 8 | 2 | .5 | .6 | 100.0 |
| KEINE ANGABE | 0 | 80 | 20.2 | MISSING | |
| | | 397 | 100.0 | 100.0 | |

b) Ordinalvariable: Schulbildung (Reichstagskandidaten 1912)

| VALUE LABEL | VALUE | FREQUENCY | PERCENT | PERCENT | PERCENT |
|---|---|---|---|---|---|
| VOLKSSCHULE | 1 | 253 | 63.7 | 74.2 | 74.2 |
| MITTELSCH O ABSCHL | 2 | 35 | 8.8 | 10.3 | 84.5 |
| HOEH SCH O EINJAEHR | 4 | 10 | 2.5 | 2.9 | 87.4 |
| HOEH SCH M EINJAHR | 5 | 1 | .3 | .3 | 87.7 |
| HOEH SCH M ABITUR | 6 | 1 | .3 | .3 | 88.0 |
| LEHRERSEM M ABSCHL | 7 | 6 | 1.5 | 1.8 | 89.7 |
| UNIV O ABSCHLUSS | 8 | 8 | 2.0 | 2.3 | 92.1 |
| UNIV M ABSCHLUSS | 9 | 27 | 6.8 | 7.9 | 100.0 |
| KEINE ANGABE | 0 | 56 | 14.1 | MISSING | |
| | TOTAL | 397 | 100.0 | 100.0 | |

*Abb. 2.3 b* zeigt die Häufigkeitsverteilung der Schulbildung der SPD-Kandidaten, die als Ordinalvariable mit 8 Kategorien definiert wurde. Der dritte Rang (»Höhere Schule ohne Einjähriges«) ist hier willkürlich nicht mit der Ziffer »3«, sondern mit der Ziffer »4« kodiert worden. Solange die Schulbildung nicht als metrische Variable, sondern als Rangskala behandelt wird, spielt das keine Rolle.

Als Beispiel für eine Intervallskala dient uns das Jahr der beruflichen Ersteinstellung (*Abb. 2.3 c*)

Für die Darstellung in der Häufigkeitstabelle sind in der Variablen BERUF1 jeweils zehn Jahre zusammengefaßt und mit einer fortlaufenden Codeziffer versehen worden. Der entsprechende SPSS$^x$-Befehl hierzu lautet

RECODE BERUF1 (1869 THRU 1878 = 1) (1879 THRU 1888 = 2) usw.

Für die Bildung von Werte-Klassen gibt es zwar einige Faustregeln, aber je nach Datenlage und Fragestellung weicht man von ihnen ab. Sie seien trotzdem hier genannt:

(1) Man bilde möglichst Klassen mit gleicher Intervallbreite
(2) Man vermeide nach Möglichkeit offene Klassen, also Klassen bei denen eine (untere oder obere) Klassengrenze nicht spezifiziert ist.
(3) Die Klassengrenzen sollten nach Möglichkeit nicht oder nur schwach besetzt sein.

Die Klassenbildung erhöht die Übersichtlichkeit; es gehen andererseits auch Informationen verloren. Deshalb kann man als »Oberregel« formulieren: Wähle die Klasseneinteilung so, daß das Charakteristische der originären Häufigkeitsverteilung (ohne Klassenbildung) erhalten bleibt. Das bedeutet z. B., daß eine etwaige Massierung der Fälle am Anfang, in der Mitte oder am Ende der nicht klassierten Skala auch nach der Klassierung in dieser Region erkennbar sein muß. Demonstrationsbeispiele hierfür bietet J. Kriz (1973: 43 ff.).

An diesem Beispiel können wir nun auch das Konzept der **kumulierten Häufigkeiten** erläutern, die in prozentuierter Form in der letzten Spalte der Häufigkeitstabelle aufgelistet sind. Wie schon erwähnt, sind sie für Ordinaldaten und höhere Meßniveaus sinnvoll interpretierbar. Um sie zu errechnen, müssen die m Ränge bzw. Werte oder Werteklassen zunächst ihrer Größe nach geordnet sein. Sodann zählt man zu den Häufigkeiten der einzelnen Ränge, Werte oder Klassen die Häufigkeiten aller davorliegenden (rangniederen) Werte (Ausprägungen) hinzu. Die bis zur Klasse oder Ausprägung $x_j$ ($j = 1,2,...,m$) kumulierte Häufigkeit $f_{cum}(x_j)$ läßt sich formal definieren durch

c) Intervallskala: Jahr der Ersteinstellung (Reichstagskandidaten 1912)

| VALUE LABEL | VALUE | FREQUENCY | PERCENT | VALID PERCENT | CUM PERCENT |
|---|---|---|---|---|---|
| 1869 BIS 1878 | 1 | 9 | 2.3 | 3.0 | 3.0 |
| 1879 BIS 1888 | 2 | 12 | 3.0 | 3.9 | 6.9 |
| 1889 BIS 1898 | 3 | 90 | 22.7 | 29.5 | 36.4 |
| 1899 BIS 1908 | 4 | 167 | 42.1 | 54.8 | 91.1 |
| 1909 BIS 1918 | 5 | 27 | 6.8 | 8.9 | 100.0 |
| KEINE ANGABE | 0 | 92 | 23.2 | MISSING | |
| | TOTAL | 397 | 100.0 | 100.0 | |

d) Ratioskala: Lebensalter (Reichstagskandidaten 1912)

| VALUE LABEL | VALUE | FREQUENCY | PERCENT | VALID PERCENT | CUM PERCENT |
|---|---|---|---|---|---|
| 25 BIS 29 | 1 | 3 | .8 | .8 | .8 |
| 30 BIS 34 | 2 | 28 | 7.1 | 7.3 | 8.1 |
| 35 BIS 39 | 3 | 71 | 17.9 | 18.6 | 26.8 |
| 40 BIS 44 | 4 | 89 | 22.4 | 23.4 | 50.1 |
| 45 BIS 49 | 5 | 85 | 21.4 | 22.3 | 72.4 |
| 50 BIS 54 | 6 | 45 | 11.3 | 11.8 | 84.3 |
| 55 BIS 59 | 7 | 32 | 8.1 | 8.4 | 92.7 |
| 60 BIS 64 | 8 | 12 | 3.0 | 3.1 | 95.8 |
| 65 BIS 69 | 9 | 8 | 2.0 | 2.1 | 97.9 |
| 70 U. AELTER | 10 | 8 | 2.0 | 2.1 | 100.0 |
| KEINE ANGABE | 0 | 16 | 4.0 | MISSING | |
| | TOTAL | 397 | 100.0 | 100.0 | |

e) Dichotome Variable: Ausübung des erlernten Berufes (Reichstagskandidaten 1912)

| VALUE LABEL | VALUE | FREQUENCY | PERCENT | VALID PERCENT | CUM PERCENT |
|---|---|---|---|---|---|
| UEBT IHN NOCH AUS | 1 | 49 | 12.3 | 12.3 | 12.3 |
| UEBT IHN NICHT MEHR | 2 | 348 | 87.7 | 87.7 | 100.0 |
| | TOTAL | 397 | 100.0 | 100.0 | |

$$(2\text{-}4) \qquad f_{cum}(x_j) = \sum_{i=1}^{j} f(x_i) \;,\; j \leq m$$

In unserem Beispiel in *Abb. 2.3 c* ist m = 5. Wählen wir j = 3, so erfahren wir, daß von den SPD-Kandidaten der Wahl von 1912, für die entsprechende Angaben vorliegen, 36.4 % ihre berufliche Ersteinstellung bis spätestens 1898 gefunden hatten.

Auf ähnliche Weise erfahren wir in der nächsten Häufigkeitstabelle (*Abb. 2.3 d*), daß 26.8 % der SPD-Kandidaten bei dieser Wahl höchstens 39 Jahre alt waren.

Wenn man Operationen gemäß (2-4) durchführt und dabei den Index »j« von 1 bis m schrittweise erhöht, erhält man eine Folge von Größen, wie sie in der letzten Spalte der Häufigkeitstabelle aufgelistet sind. Trägt man die Folge von kumulierten relativen Häufigkeiten auf der Ordinate und die entsprechenden $x_j$-Werte (j = 1,2,...,m) auf der Abszisse eines Koordinatenkreuzes ein, erhält man die **empirische Verteilungsfunktion** $F(x) = f_r(X \leq x)$. Verteilungsfunktionen spielen in der Inferenzstatistik eine wichtige Rolle. Wir werden sie in Kap. 6.5 eingehender besprechen.

## 2.3 Graphische Darstellung von Häufigkeitsverteilungen

Mehr noch als die Häufigkeitstabellen vermitteln graphische Darstellungen einen unmittelbaren Eindruck von den charakteristischen Merkmalen einer Häufigkeitsverteilung. Je nach Meßniveau gibt es hierzu unterschiedliche Möglichkeiten.

Häufigkeitsverteilungen metrischer Daten stellt man in der Regel in Form eines **Histogramms** dar. Dabei werden die Häufigkeiten auf den positiven Achsen eines Koordinatenkreuzes eingetragen: die Variablenwerte bzw. Klassen auf der x-Achse (Abszisse), die Häufigkeiten auf der y-Achse (Ordinate). Von den Schnittpunkten der Koordinaten werden Linien oder Säulen (Balken) auf die Variablenwerte gezogen. Die Säulenhöhen sind also proportional den Häufigkeiten. (In SPSS$^x$ wird zwischen dem Balkendiagramm BARCHART und dem Liniendiagramm HISTOGRAM unterschieden. Damit werden aber lediglich unterschiedliche graphische Aufmachungen für die gleiche Sache bezeichnet.)

Verbindet man die Mittelpunkte der Säulenenden mit Linien, erhält man den sog. **Polygonzug** (*Abb. 2.5*, diesmal in der »korrekten« Form: ohne Lücken zwischen den Säulen).

Die Darstellungsform des Histogramms benutzt man häufig auch bei Ordinal und Nominalskalen. Man läßt dann aber Lücken zwischen den

Abb. 2.4: Histogramm der gruppierten Altersverteilung (siehe Abb. 2.3.d)

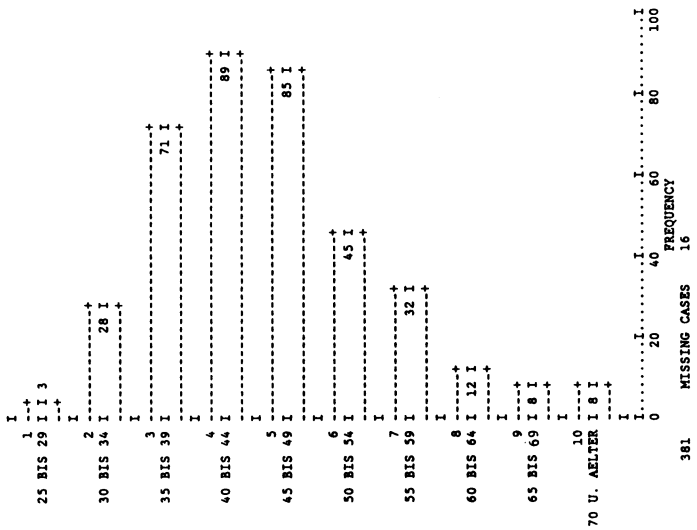



Abb. 2.5: Polygonzug zum Histogramm in Abb. 2.4

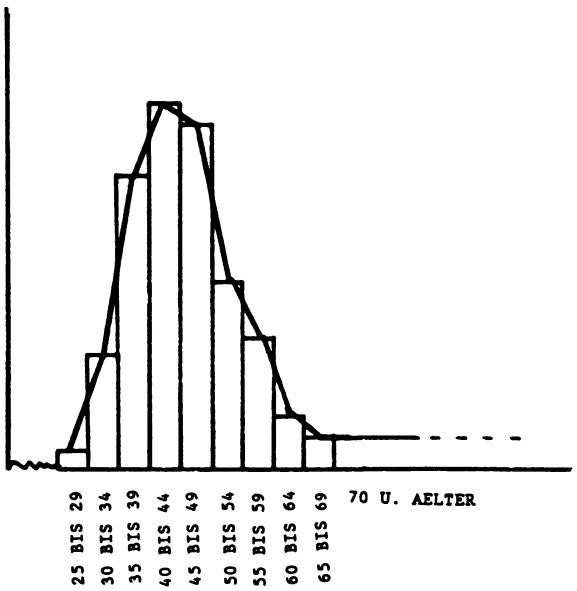


Abb. 2.6: Histogramm zur Häufigkeitsverteilung in Abb. 2.3.b nach Zusammenfassung von Kategorien (Schulbildung)

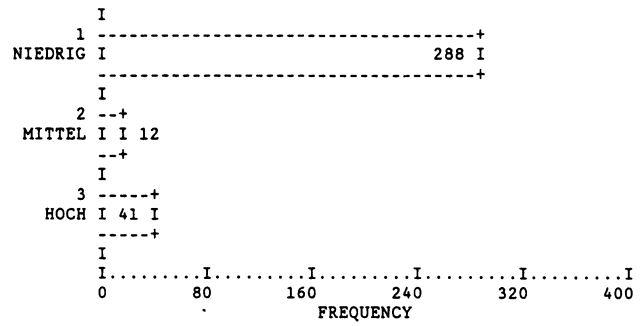
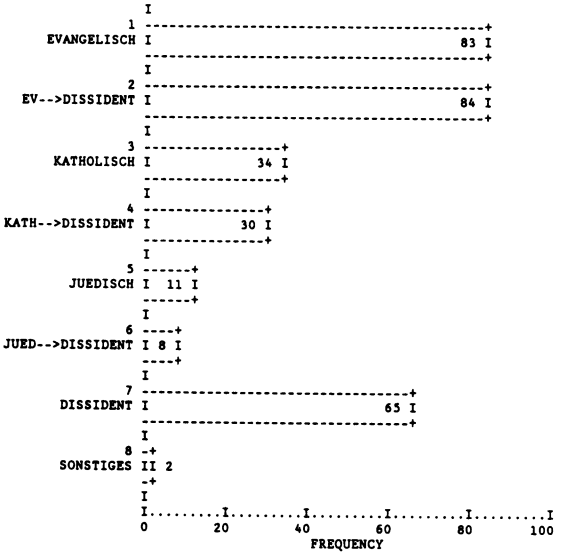



Abb. 2.7: Histogramm zur Häufigkeitsverteilung in Abb. 2.3a (Konfession)



Säulen (oder »Stäben«), um nicht den falschen Eindruck aufkommen zu lassen, es liege ein Wertekontinuum vor. Das BARCHART-Subkommando in SPSS$^{x}$ richtet diese Lücken automatisch ein (allerdings auch bei metrischen Daten). Bei ordinalen Daten ist die Reihenfolge der Kategorien (Rangplätze) zu beachten; bei nominalen Daten spielt sie natürlich keine Rolle. Dazu wieder zwei Beispiele aus unserem Datensatz (*Abb. 2.6, Abb. 2.7*).

Liegen nur wenige Merkmalsausprägungen bzw. Kategorien vor (wie meistens bei Nominal- und Ordinalskalen), sind auch Kreis- und Streifendiagramme häufig gewählte Darstellungsformen (*Abb. 2.8, Abb. 2.9*).

Graphische Darstellungen können leicht für manipulative Zwecke mißbraucht werden. Kriz (1973: 48 ff.) erläutert einige Beispiele.

In der Literatur werden verschiedene Verteilungsformen unterschieden, die wir hier aus dem Lehrbuch von Jürgen Bortz (1979) reproduzieren (*Abb. 2.10*). Einstweilen benötigen wir sie lediglich zur terminologischen Verständigung.

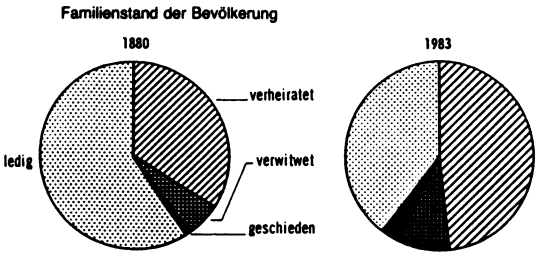
Die in der Abbildung als rechts- und linkssteil bezeichneten Verteilungsformen werden andernorts in der Literatur auch als links- und rechtsschief bezeichnet (rechtssteil = linksschief).

## 2.4 Häufigkeitsdichte

Histogramm-Darstellungen können bei metrischen Daten mißverständlich werden, sobald die auf der x-Achse eingetragenen Intervalle, also die Grundlinien der Säulen, unterschiedlich lang sind. Wird nämlich die Säulenhöhe weiterhin proportional zu den Häufigkeiten gehalten, müssen die von den Säulenkanten eingeschlossenen Flächen zu ihnen **dis**proportional sein. Ein doppelt so langes Intervall z. B. verdoppelt bei gleichbleibender Häufigkeit (gleicher Säulenhöhe) die eingeschlossene Fläche. *Abb. 2.11* veranschaulicht diesen Vorgang.

Die größere Fläche suggeriert unwillkürlich eine doppelte Häufigkeit, da sich das Auge nun einmal nicht nur an Höhen (bzw. Längen), sondern auch an Flächen orientiert. Tatsächlich würde aber die gleiche Häufigkeit bei doppelt so großem Intervall bedeuten, daß die Fälle in diesem Wertebereich »dünner« gestreut sind, daß die Fläche gleichsam weniger »dicht besiedelt« ist. Das bringt uns zu der Idee, nicht die Höhen, sondern die Flächenprojektionen der Säulen proportional zu den Häufigkeiten zu gestalten und die Höhen entsprechend anzupassen. Die folgende *Abb. 2.12* »korrigiert« in diesem Sinne die *Abb. 2.11*.

Abb. 2.8: Kreisdiagramme




Quelle: Datenreport 1985, S. 47

Abb. 2.9: Streifendiagramme

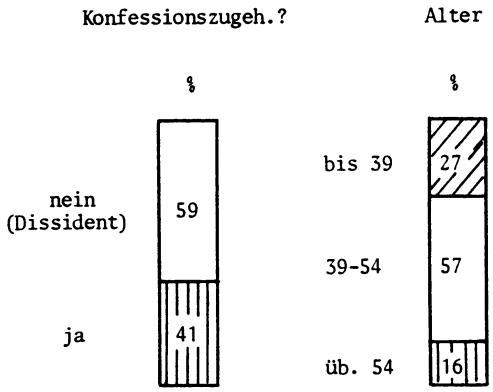
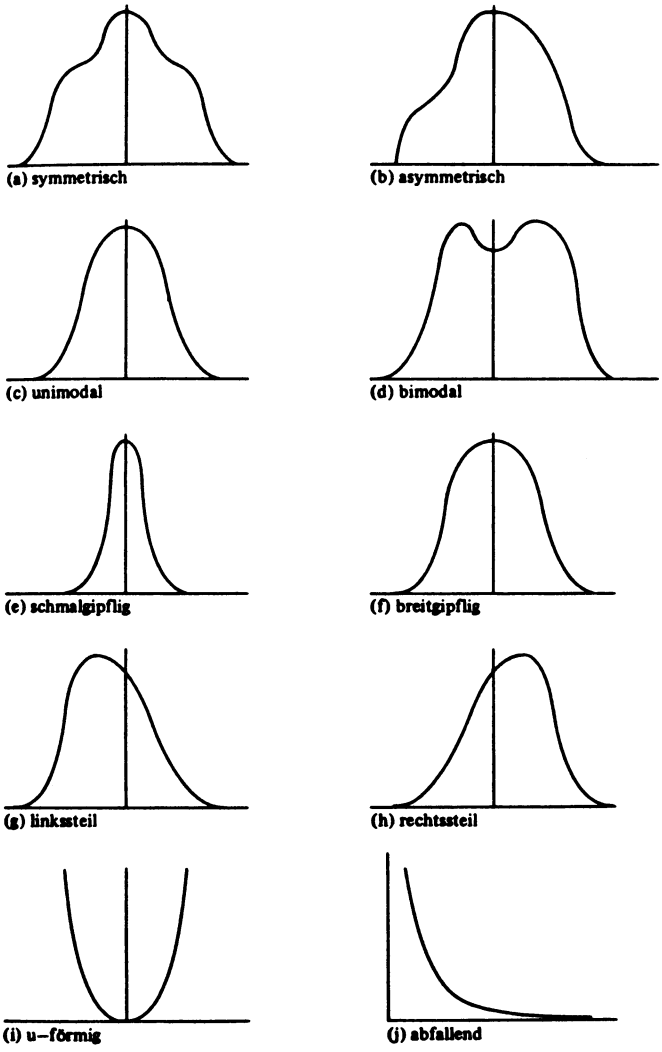
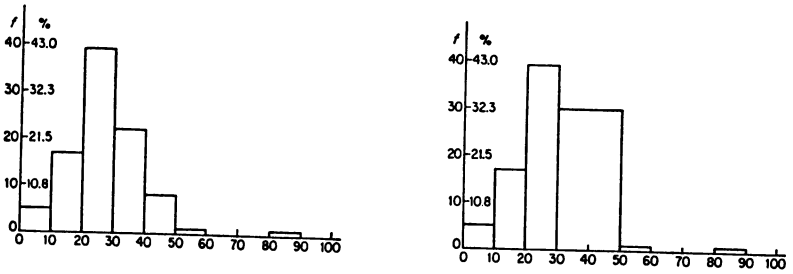


Abb. 2.10: Formen univariater Verteilungen



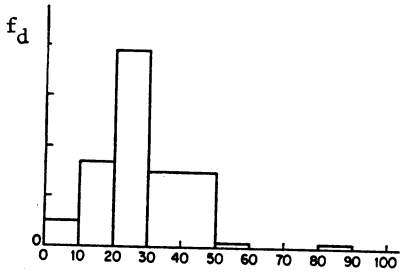

Quelle: Bortz 1979, S. 41

Abb. 2.11: Histogramme mit gleichen und ungleichen Intervallen




Quelle: Blalock 1960, S. 40. f

Abb. 2.12: Relative Häufigkeitsdichten
(Flächensegmente proportional zur Häufigkeit)




Quelle: Blalock 1960, S. 41 (modifiziert)

Numerisch erreichen wir dies, indem wir die relativen Häufigkeiten durch die Intervallängen (Klassenbreiten) dividieren. Dadurch erhalten wir als neue Größe die

$$\text{(2-5)} \qquad \text{Relative Häufigkeitsdichte} = \frac{\text{Relative Häufigkeit}}{\text{Klassenbreite}}$$

Auf der Ordinate werden nun nicht mehr die Häufigkeits-, sondern die Dichtewerte $f_d$ eingetragen. So wie zuvor die relativen Häufigkeiten sich zum Betrag »1« summierten, summieren sich auch die Flächensegmente über den Intervallen zum Betrag »1« (Einheitsfläche). Das ergibt sich aus der Umformung der Definitionsgleichung (2-5):

$$\text{(2-6)} \qquad \begin{aligned}\text{Rel. Häufigkeit} &= \text{rel. Häufigk.dichte} \cdot \text{Klassenbreite} \\ &= \text{Säulenfläche}\end{aligned}$$

Wenn man über alle Flächensegmente auf der rechten Seite der Gleichung summiert, muß sich die Größe »1« ergeben, da die Summe der relativen Häufigkeiten (linke Seite der Gleichung) gleich 1 ist.

Bei einer kontinuierlichen Variablen lassen sich die Klassenintervalle beliebig verkleinern. Die oberen Säulenkanten nähern sich dann der Form einer stetigen Kurve.

Der Begriff der Häufigkeitsdichte wird zwar hier im Rahmen der deskriptiven Statistik vorgestellt. Seine wichtigere Bedeutung erhält er aber dadurch, daß er ein zentrales Konzept der Inferenzstatistik, nämlich das der Wahrscheinlichkeitsdichte (siehe Kap. 6) vorbereitet.

## 2.5 Exkurs: Zusammenlegen von Kategorien und Variablen

Gelegentlich möchte der Forscher zwei (oder mehr) ursprünglich getrennte Variablen zu einer neuen Variablen zusammenfassen. So sind z. B. die früheren politischen Erfahrungen von Abgeordneten der Frankfurter Nationalversammlung u. a. in zwei Variablen erfaßt worden, deren univariate Verteilungen in den *Abb. 2.13 a,b* wiedergegeben sind:

Wir wollen nun beide Variablen vereinfachen und dann zu einer neuen Variablen »Politische Erfahrungen vor 1848« (POLERF) zusammenfassen. Sie soll folgende Kategorien enthalten:

1 Vor 1848 ausschließlich in politischen Ämtern tätig gewesen
2 Sowohl Amtserfahrung als auch illegale bzw. oppositionelle Aktivitäten vor 1848 (»Inkonsistente«)
3 Ausschließlich illegale bzw. oppositionelle Aktivitäten

0 Weder oppositionell/illegal noch als Amtsträger vor 1848 tätig gewesen

Dazu legen wir zunächst die Kategorien 1, 2, 3 der INSTAKT-Variablen zusammen, die danach nur noch zwei Ausprägungen enthält: keine politischen Ämter (0), politische Ämter (1). Ähnlich verfahren wir mit der Variablen NINSTAKT, in der wir politisch oppositionelle Tätigkeiten von den anderen Aktivitäten abgrenzen. Die entsprechenden RECODE-Befehle für beide Variablen lauten:

```
RECODE INSTAKT (1, 2, 3 = 1)
       NINSTAKT (1 = 0) (2,3 = 1)
```

Abb. 2.13a: Institutionalisierte politische Erfahrungen vor 1848 (Variable INSTAKT)

| VALUE LABEL | VALUE | FREQUENCY | PERCENT | VALID PERCENT | CUM PERCENT |
|---|---|---|---|---|---|
| NICHT ZUTREFFEND | 0 | 569 | 70.3 | 70.3 | 70.3 |
| GEMEINDEAMT | 1 | 53 | 6.6 | 6.6 | 76.9 |
| REG+STAEND VERTR | 2 | 30 | 3.7 | 3.7 | 80.6 |
| PARL+STAAT | 3 | 157 | 19.4 | 19.4 | 100.0 |
| | TOTAL | 809 | 100.0 | 100.0 | |

VALID CASES 809 MISSING CASES 0

Abb. 2.13b: Nichtinstitutionalisierte politische Erfahrungen vor 1848 (Variable NINSTAKT)

| VALUE LABEL | VALUE | FREQUENCY | PERCENT | VALID PERCENT | CUM PERCENT |
|---|---|---|---|---|---|
| NICHT ZUTREFFEND | 0 | 527 | 65.1 | 65.1 | 65.1 |
| KOMPENS PARTIZ | 1 | 63 | 7.8 | 7.8 | 72.9 |
| POL PUBLIZ | 2 | 86 | 10.6 | 10.6 | 83.6 |
| POL VEREIN | 3 | 133 | 16.4 | 16.4 | 100.0 |
| | TOTAL | 809 | 100.0 | 100.0 | |

VALID CASES 809 MISSING CASES 0

Die rekodierten Variablen erlauben 2x2 = 4 Kombinationen, die die Ausprägungen der zusammengefaßten Variablen bilden (s. *Abb. 2.14*).

Abb.2.14: Kombinationsmöglichkeiten der dichotomisierten Variablen zur politischen Erfahrung

| | | NINSTAKT | | |
|---|---|---|---|---|
| | COUNT | I | | |
| | ROW PCT | Ikeine op politisc | | ROW |
| | COL PCT | Ip. pol. | he Erf. | TOTAL |
| | | I 0 | I 1 | I |
| INSTAKT | | +---------+---------+ | | |
| 0 | | I 433 | I 136 | I 569 |
| keine Amtserfahr | | I 76.1 | I 23.9 | I 70.3 |
| | | I 73.4 | I 62.1 | I |
| | | +---------+---------+ | | |
| 1 | | I 157 | I 83 | I 240 |
| Amtserfahrung | | I 65.4 | I 34.6 | I 29.7 |
| | | I 26.6 | I 37.9 | I |
| | | +---------+---------+ | | |
| | COLUMN | 590 | 219 | 809 |
| | TOTAL | 72.9 | 27.1 | 100.0 |

Abb. 2.15: Politische Erfahrungen der Abgeordneten der Frankfurter Nationalversammlung (Varibale POLERF)

| VALUE LABEL | VALUE | FREQUENCY | PERCENT | VALID PERCENT | CUM PERCENT |
|---|---|---|---|---|---|
| K. POLIT. ERFAHRUNG | 0 | 433 | 53.5 | 53.5 | 53.5 |
| NUR AMTSERFAHRUNG | 1 | 157 | 19.4 | 19.4 | 72.9 |
| INKONSISTENTE ERFAHR | 2 | 83 | 10.3 | 10.3 | 83.2 |
| NUR OPPOSITION | 3 | 136 | 16.8 | 16.8 | 100.0 |
| | TOTAL | 809 | 100.0 | 100.0 | |

VALID CASES 809 MISSING CASES 0

Die neue Variable POLERF wird in SPSS$^x$ mit dem COMPUTE-Befehl definiert. Um diese Variable einzurichten, empfiehlt sich zunächst ein Kommando, das allen Abgeordneten (allen Fällen) auf dieser Variable den Wert 0 zuweist:

```
COMPUTE    POLERF = 0
```

Für alle Abgeordnete, die diesen Wert nicht behalten sollen, werden neue Werte gemäß der Tabelle in *Abb. 2.14* über IF-Statements zugewiesen:

```
IF(INSTAKT EQ 1 AND NINSTAKT EQ 0)POLERF = 1
IF(INSTAKT EQ 1 AND NINSTAKT EQ 1)POLERF = 2
IF(INSTAKT EQ 0 AND NINSTAKT EQ 1)POLERF = 3
```

Da in beiden Variablen keine Werte fehlen, müssen alle Fälle, für die keiner der logischen Ausdrücke in den drei IF-Statements zutrifft, die Kombination INSTAKT EQ 0 AND NINSTAKT EQ 0 aufweisen. Über das COMPUTE-Statement ist diesen Fällen aber bereits der Wert 0 für die POLERF-Variable zugewiesen worden (für alle anderen Fälle ist dieser Wert durch die IF-Statements korrigiert worden). Auf die neu gebildete Variable POLERF werden wir in Kap. 5 zurückkommen. Ihre univariate Verteilung ist in *Abb. 2.15* widergegeben.

Das Zusammenfügen von Variablen ist etwas komplizierter, wenn fehlende Werte beachtet werden. Nehmen wir einmal an, daß bei einigen Abgeordneten für die Variablen INSTAKT und NINSTAKT keine validen Angaben vorliegen. In diesem Falle müssen die fehlenden Werte zunächst einmal als solche kodiert und deklariert werden. Wird ihnen z. B. der Wert 9 zugewiesen, lautet das entsprechende SPSS$^x$-Kommando:

```
MISSING VALUES INSTAKT,NINSTAKT(9)
```

In diesem Falle wird die neu zu bildende Variable (in unserem Beispiel POLERF) besser nicht mit COMPUTE = 0, sondern mit

```
NUMERIC POLERF(F1)
```

initialisiert. Damit werden **alle** Fälle zunächst auf den in SPSS$^x$ intern eingerichteten Missing-Value-Code gesetzt, der im Computer-Ausdruck als Punkt (.) erscheint. Der Klammerzusatz F1 besagt lediglich, daß POLERF als einstellige Ziffer ohne Komma gebildet werden soll. Wie vorher der durch COMPUTE erzeugte Wert Null, kann nun auch der Missing-Value-Code durch IF-Statements für diejenigen Fälle korrigiert werden, die die entsprechende logische Bedingung erfüllen. Diesmal benötigen wir aber vier, nicht drei IF-Statements, um alle Möglichkeiten der Kombination von INSTAKT- und NINSTAKT-Werten zu erfassen und POLERF zuzuweisen.

Beim Gebrauch aufeinanderfolgender IF-Statements in SPSS$^x$ ist allgemein zu beachten, daß sie eine hierarchisch aufsteigende Ordnung bilden, so daß ein nachfolgendes IF ein vorausgegangenes IF in der Ergebnisvariable überschreiben kann. Eine alternative, nicht-hierarchische Folge von IF-Statements erhält man z. B. durch folgende (hier vereinfachte) Struktur:

```
DO IF (logische Bedingung)
Anweisungen, z. B. POLERF = 0
ELSE IF (logische Bedingung)
Anweisungen, z. B. POLERF = 1
END IF
```

Nähere Erläuterungen hierzu sind den einschlägigen Handbüchern zum Programmsystem SPSS$^x$ zu entnehmen.

# Kapitel 3

## Maßzahlen zur Kennzeichnung univariater Verteilungen

Nachdem wir Häufigkeitsverteilungen graphisch dargestellt haben, können wir fragen, ob sich deren charakteristische Merkmale auch numerisch ausdrücken lassen. Damit wäre zwar ein Informationsverlust verbunden; aber Maßzahlen sind leichter vergleichbar und eher mitteilbar als Datenmatrizen, Häufigkeitstabellen und Abbildungen. Außerdem ist es gerade die Aufgabe eines Wissenschaftlers, aus einer Vielzahl von Informationen das »Wesentliche« herauszugreifen und zusammenzufassen.

Häufigkeitsverteilungen lassen sich zunächst einmal nach ihren Mittelwerten und nach dem Ausmaß der Streuung, der Variabilität der einzelnen Werte kennzeichnen. So unterscheidet man zwischen den Maßzahlen der »zentralen Tendenz« (Lokalisations- oder Lokationsmaße, Lageparameter) und den Streuungs- oder Dispersionsmaßen. Je nach Meßniveau sind hierfür unterschiedliche Kennzahlen definiert. Die gebräuchlisten wollen wir nun vorstellen:

### 3.1 Lokalisationsmaße

**Der Modus (h):**

Der Modus (Modalwert, dichtester Wert) ist definiert als der am häufigsten vorkommende Wert einer Verteilung. In unserem Beispiel aus *Abb. 2.3 a* [Nominalskala] ist das der Wert h = 2 (ev., Dissident), im Beispiel aus *Abb. 2.3 b* [Ordinalskala] ist h = 1 (Volksschule). Bei gruppierten (klassierten) Werten wird die Mitte derjenigen Klasse, die die größte Häufigkeit aufweist, als Modalwert betrachtet. In unserem Beispiel aus *Abb. 2.3 c* [Intervallskala] ist also h = 4 bzw. 1903,5.

Wenn nebeneinander liegende Werte gleich häufig vorkommen und ihre Häufigkeit größer als die benachbarter Werte ist, so ist das arithmetische Mittel der häufigsten Werte als Modus definiert. Im folgenden Beispiel (*Abb. 3.1*) ist h = 5,5.

Wenn eine Verteilung mehrere »Gipfel« also mehrere **relative** Häufigkeitsmaxima aufweist, die nicht unmittelbar benachbart sind, spricht man von bi- oder multimodalen Verteilungen - wie in *Abb. 3.2*.

In einem solchen Falle gibt man für jedes relative Häufigkeitsmaximum einen Modalwert an, hier also $h_1 = 5$ und $h_2 = 8$.

Der Modus ist das einzige Lokalisationsmaß, das auch für Nominalskalen definiert ist. Alle folgenden Lageparameter setzen ein höheres Meßniveau voraus.

Abb. 3.1

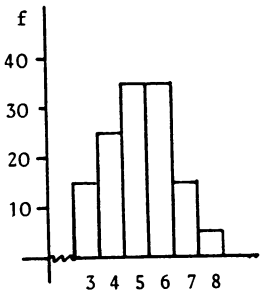
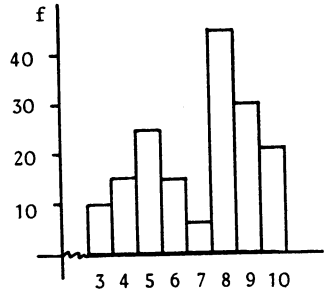



Abb. 3.2




**Der Median ($\tilde{x}$):**

Der Median setzt mindestens ordinales Meßniveau voraus. Die Werte müssen ihrer Größe nach geordnet sein. Dann läßt sich der Median ($\tilde{x}$) definieren, als »derjenige Wert, der genauso viele kleinere (oder gleichgroße) Werte vor sich wie größere (oder gleichgroße) hinter sich hat« (Kriz 1973: 56). Mit anderen Worten, der Median halbiert eine nach ihrer Größe geordnete Reihe von Meßwerten.

Wenn die Anzahl der Fälle (n) ungerade ist, ist der Wert des mittleren, $(n+1)/2$-ten Falles der Median. Die folgende Rangreihe umfaßt $n = 7$ Fälle:

5 6 7 7 8 10 10

Da $(7+1)/2 = 4$, ist der Wert des vierten Falles, $x_4 = 7$, gleich dem Median $\tilde{x}$.

Bei einer geraden Anzahl von Fällen und metrischen Daten betrachtet man die halbierte Wertesumme der beiden mittleren Fälle als Median. Gegeben seien zum Beispiel die folgenden 8 Werte:

5 6 7 7 8 10 10 11

Die beiden mittleren Fälle haben die Werte 7 und 8. Folglich ist der Median $\tilde{x} = (7+8)/2 = 7{,}5$. Falls es sich bei dieser Wertereihe nicht um eine metrische, sondern um eine Rangskala handelt, sollte man sich mit der Aussage begnügen: »Der Median liegt zwischen den Werten 7 und 8«.

Bei gruppierten Daten erhält man ein »Medianintervall«. Als solches ist dasjenige Intervall definiert, in dem bei einer geordneten Meßreihe der

n/2-te Fall liegt. Wenn z. B. 110 Werte vorliegen und nach Größe geordnet sind, ist dies das Intervall des 55. Falls. Auch in einer solchen Situation kann man bei metrischen Daten einen einzelnen Medianwert durch Interpolation berechnen. Die Formel hierfür ist z. B. in Benninghaus (1976, S. 41 f.) erläutert. Da wir uns diesen Wert jederzeit vom Computer ausdrucken lassen können, verzichten wir darauf, sie hier zu erläutern.

Der Median ist (anders als das arithmetische Mittel, siehe unten) gegenüber einzelnen »Ausreißern« (extrem vom Durchschnitt abweichenden Werten) unempfindlich. Man bezeichnet ihn deshalb auch als »robusten« Lageparameter.

Er hat zudem die mathematische Eigenschaft, die Summe $\Sigma \mid x_i - \tilde{x} \mid$ zu einem Minimum zu machen. Das heißt, wenn man für $\tilde{x}$ irgendeinen anderen positiven Betrag einsetzt (z. B. das arithmetische Mittel), wird die Summe der absoluten Abweichungsbeträge größer. (Den Beweis hierfür findet man z. B. in Schlittgen 1987: 114). Ein Beispiel zur Illustration: Gegeben seien die Werte 1,2,3. In dieser Verteilung ist $\tilde{x} = 2$, also gleich dem arithmetischen Mittel. Nehmen wir weiter an, infolge eines Kodierfehlers sei statt der 3 eine 33 eingetragen worden. Dann ist der Median unverändert $\tilde{x} = 2$, das arithmetische Mittel aber $\bar{x} = 12$. Die Summe der absoluten Abweichungsbeträge ist, bezogen auf den Median, gleich 32, bezogen auf das arithmetische Mittel: 42.

**Das arithmetische Mittel: ($\bar{x}$)**

Es ist definiert als Summe der Meßwerte, dividiert durch ihre Anzahl:

$$(3\text{-}1) \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Seine Berechnung setzt metrische Daten voraus. Kommen Meßwerte mehr als einmal vor, gibt es also nur $k < n$ verschiedene Werte, müssen sie nicht einzeln addiert, sondern können mit ihrer jeweiligen Häufigkeit $f_i$ multipliziert werden, mit der sie auftreten:

$$(3\text{-}2) \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{k} x_i \cdot f_i \;, \; k < n$$

Bei gruppierten Daten werden die Klassenmitten als Meßwerte $x_i$ betrachtet. Man sollte das arithmetische Mittel aber nach Möglichkeit aus den ungruppierten Daten berechnen, da man nicht voraussetzen kann, daß die Klassenmitte auch der Mittelwert aller Werte innerhalb der Klasse ist.

Das arithmetische Mittel hat folgende mathematische Eigenschaften, von denen man z. B. bei Skalentransformationen Gebrauch machen kann:

(a) Addiert man eine Konstante $c \neq 0$ zu allen Werten einer Verteilung, vergrößert (bzw. verkleinert) sich das arithmetische Mittel um diesen Betrag:

Aus $x' = (x + c)$ folgt: $\overline{x'} = \bar{x} + c$

(b) Multipliziert man jeden Wert einer Verteilung mit dem Faktor $b \neq 0$, so vervielfacht sich das arithmetische Mittel um den gleichen Faktor:

Aus $x' = bx$ folgt: $\overline{x'} = b\bar{x}$

(c) Somit gilt auch: Aus $x' = c + b\bar{x}$, folgt $\overline{x'} = c + b\bar{x}$.

(d) Die Summe der Abweichungen aller Meßwerte von ihrem arithmetischen Mittel ist Null:

$$\Sigma (x_i - \bar{x}) = 0$$

(e) Die Summe der **quadrierten** Abweichungen aller Meßwerte von ihrem arithmetischen Mittel (man bezeichnet sie auch als »Variation«) ist ein Minimum. Das heißt, sie ist kleiner als die Summe der quadrierten Abweichungen aller Meßwerte von einer anderen Konstanten $c \neq \bar{x}$:

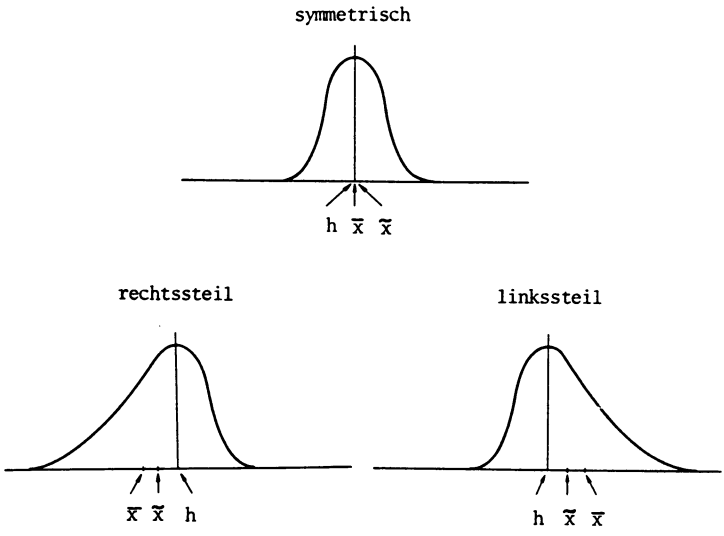
$$\Sigma (x_i - \bar{x})^2 = \min.$$

Das ist eine Eigenschaft, die die »Methode der Kleinstquadrate« bei der Bestimmung der Regressionskoeffizienten (siehe Kap. 10) mit begründet.

Das arithmetische Mittel kennzeichnet eine Verteilung um so unvollkommener, je asymmetrischer (»schiefer«) sie ist. Einzelne »Ausreißer« (wenige extrem hohe oder niedrige Werte) verschieben das arithmetische Mittel leicht nach oben oder unten, so daß die »zentrale Tendenz« der Verteilung in diesem Falle eher durch den Median auszudrücken wäre. Wenig aussagekräftig ist das arithmetische Mittel auch dann, wenn die Verteilung mehrgipflig ist.

Ist eine Verteilung sowohl symmetrisch als auch unimodal, fallen Modus, Median und arithmetisches Mittel zusammen. Bei links- und rechtssteilen Verteilungen (siehe oben, *Abb. 2.10*) ergeben sich umgekehrte Rangfolgen für diese Maßzahlen - wie die folgende *Abb. 3.3* zeigt. Aus der Größenordnung der Differenzen erhält man also Hinweise über den Grad der Schiefe.

Abb. 3.3: Position unterschiedlicher Mittelwerte in verschiedenen Verteilungen




Quelle: Bortz 1979, S. 49 (modifiziert)

### 3.2 Streuungsmaße

Wie wir sahen, kennzeichnen die Lokalisationsmaße allein eine Verteilung nur sehr unvollkommen. Bei gleichem arithmetischen Mittel z. B. können die einzelnen Werte sehr unterschiedlich streuen. Deshalb muß mindestens noch ein Streuungsmaß als zusätzliche Kennzahl mit angegeben werden. Hierzu sind verschiedene Konzepte vorgeschlagen worden, von denen wir nun einige erläutern wollen.

**Spannweite (Range):**

Das vielleicht naheliegendste Streuungsmaß ist die Spannweite bzw. der »Range« (R), der als die Differenz zwischen dem größten und dem kleinsten Meßwert einer Verteilung definiert ist:

$$(3\text{-}3) \qquad R = x_{max} - x_{min}$$

Bei gruppierten Daten wird die Differenz zwischen den Mittelpunkten der extremen Klassenintervalle angegeben. Die Spannweite ist als Kenngröße nur für metrische Daten sinnvoll. Sie ist aber auch dort wenig aussagekräftig, da völlig offen bleibt, wie die Verteilung zwischen den Extremwerten aussieht.

**Centile, Quartile, Quantile:**

Wir haben den Median als eine Maßzahl kennengelernt, die eine Verteilung mit mindestens ordinalskalierten Werten in zwei gleich große Hälften teilt. Analog dazu kann man, wenn entsprechend viele Fälle nach der Größe ihrer Meßwerte geordnet sind, eine Verteilung in hundert Abschnitte zerlegen. Man spricht dann von »Centilen« (C). Jeweils 25 Centile entsprechen einem »Quartil« (Q). $C_{25} = Q_1$ ist der Wert desjenigen Falles, bei dem das erste Viertel der Verteilung endet. $Q_2$ ist identisch mit dem Median. Als Streuungsmaße verwendet man gelegentlich den

$$(3\text{-}4) \qquad \text{Quartilabstand} = Q_3 - Q_1 = C_{75} - C_{25}$$

oder den

$$(3\text{-}5) \qquad \text{mittleren Quartilabstand} = \frac{Q_3 - Q_1}{2}$$

Wenn die Verteilung symmetrisch ist, müssen $Q_1$ und $Q_3$ gleich weit vom Median bzw. dem arithmetischen Mittel entfernt sein. Ist die Differenz zwischen dem Median und $Q_1$ größer als die Differenz zwischen $Q_3$ und dem Median, liegt eine linksschiefe Verteilung vor.

Man kann natürlich noch weitere Gruppierungen der kumulierten Häufigkeitsverteilung definieren, z. B. »Dezile« für die 10 %-, 20 %- , ... , 90 %-Punkte. Der Oberbegriff für beliebige Unterteilungen ist »Quantile«. Ohne hier eine formal exakte Definition zu geben, läßt sich ein »p-Quantil« als derjenige Wert $x_p$ eines geordneten Datensatzes $x_1,...,x_p,...x_n$ bezeichnen, bei dem die kumulierte relative Häufigkeit den Wert p erreicht. $x_{0.95}$ meint also denjenigen Wert, den nur 5 % der Fälle überschreiten. Quantilen werden wir in Teil II bei den Wahrscheinlichkeitsverteilungen und statistischen Signifikanztests begegnen. Mit Hilfe der Quantile lassen sich

in sog. **QQ-Diagrammen** zwei unterschiedliche Verteilungen auf besonders effiziente Weise graphisch miteinander vergleichen. Davon werden wir in Teil II, Kap. 10.4 Gebrauch machen.

Centile und Quartile werden gelegentlich auch für Ordinalskalen errechnet. Da Differenzen bei Ordinaldaten aber nicht quantitativ interpretierbar sind, sind Quartilabstände erst ab Intervallskalenniveau sinnvoll interpretierbar. Das gilt auch für die folgenden Streuungsmaße.

**Die durchschnittliche Abweichung:**

Sie ist definiert als arithmetisches Mittel der Absolutbeträge der Abweichungen aller Meßwerte vom arithmetischen Mittel der Verteilung. Als Abkürzung verwendet man »AD« (von der englischen Bezeichnung »average distance«).

$$(3\text{-}6) \qquad AD = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Die Absolutbeträge werden eingesetzt, weil sich sonst positive und negative Abweichungen ausgleichen und zu Null summieren könnten (siehe oben). Damit liegt ein sehr anschauliches Maß für die Streuung vor. Es ist aber in der Statistik von einer anderen Kennzahl, der Varianz bzw. Standardabweichung, weitgehend verdrängt worden. Wir werden diese »neue« Kennzahl zunächst vorstellen und dann beide Konzepte miteinander vergleichen.

**Varianz und Standardabweichung:**

Die Varianz ($s^2$) wird im Rahmen der Deskriptivstatistik definiert als Summe der **quadrierten** Abweichungen aller Meßwerte von ihrem arithmetischen Mittel, dividiert durch die Zahl n der Fälle:

$$(3\text{-}7) \qquad s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

oder bei gruppierten Daten

$$(3\text{-}7') \qquad s^2 = \frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x})^2 f_i \quad \text{bzw.} \quad \sum_{i=1}^{k} (x_i - \bar{x})^2 f_i / n$$

($x_i$ bezeichnet die Klassenmitte)

Es läßt sich beweisen (siehe z. B. Schlittgen 1987, S. 136), daß auch gilt.

$$(3\text{-}7'') \qquad s^2 = \overline{x^2} - \bar{x}^2$$

In der Inferenzstatistik wird der Begriff der (zu schätzenden) Varianz algebraisch etwas anders definiert. Es wird dann nicht durch n, sondern durch (n-1) dividiert. Um beide Kontexte zu unterscheiden, wird in der Literatur gelegentlich der Begriff der Varianz dem inferenzstatistischen Kontext vorbehalten und der Ausdruck (3-7) als »mittlere quadratische Abweichung« (»mean squared deviation«, MSD) wiedergegeben. Wir wollen hier aber den Begriff der Varianz, wie üblich, für beide Kontexte, den der beschreibenden und den der Inferenzstatistik, verwenden.

Mit dem Quadrieren der Mittelwertabweichungen erreicht man zunächst einmal, daß die Abweichungen sich nicht bei der Summierung ausgleichen. Um daraus eine lineare Maßzahl zu erhalten, zieht man aus der Varianz die Quadratwurzel. Den neuen Ausdruck bezeichnet man als Standardabweichung:

$$(3\text{-}8) \qquad s = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}$$

Im Gegensatz zur Varianz hat die Standardabweichung die gleiche Meßdimension wie die Daten, aus denen sie berechnet wurde.

Ein wesentlicher Grund, weshalb die Standardabweichung die durchschnittliche Abweichung (AD) als Maßzahl weitgehend verdrängt hat, liegt in ihrer breiteren Verwendbarkeit im Rahmen der Inferenzstatistik und des Regressionsmodells (siehe Teil II dieses Grundkurses). Als Deskriptivmaßzahlen unterscheiden sich die beiden in zwei Punkten:

(a) Durch das Quadrieren erhalten die größeren Abweichungsbeträge ein relativ größeres Gewicht in der Summenbildung. Machen wir uns das an einem kleinen Zahlenbeispiel deutlich:

Gegeben seien die Werte
2, 2, 2, 4, 4, 4
Für diese Verteilung ist $\bar{x} = 3$, AD = 1 und $s = 1$.

Betrachten wir nun eine zweite Verteilung:
2, 8, 2, 4, 4, 4
Sie unterscheidet sich von der ersten nur durch den größeren zweiten Wert. Jetzt ist $\bar{x} = 4$, AD = 1,33 und $s = 2{,}0$. Die Standardabweichung ist also wesentlich stärker angewachsen als AD.

(b) AD und s reagieren unterschiedlich auf Umverteilungen. AD reagiert nicht auf Umverteilungen, wenn sie sich innerhalb der gleichen Hälfte (ober- oder unterhalb des Mittelwertes) abspielen. Wir wollen das anhand der folgenden Verteilungen dreier Merkmalsdimensionen X, Y und Z demonstrieren:

| x | f(x) | y | f(y) | z | f(z) |
|---|---|---|---|---|---|
| 3 | 2 | 3 | 2 | 3 | 0 |
| 4 | 2 | 4 | 2 | 4 | 0 |
| 5 | 0 | 5 | 0 | 5 | 2 |
| 6 | 3 | 6 | 3 | 6 | 5 |
| 7 | 5 | 7 | 5 | 7 | 5 |
| 9 | 4 | 9 | 0 | 9 | 4 |
| 10 | 3 | 10 | 7 | 10 | 7 |
| 11 | 0 | 11 | 4 | 11 | 0 |
| 12 | 4 | 12 | 0 | 12 | 0 |
| Σ: | 23 | | 23 | | 23 |
| $\bar{x} = \bar{y} = \bar{z}$: | 7,87 | | 7,87 | | 7,87 |
| AD: | 2,39 | | 2,39 | | 1,69 |
| s: | 2,79 | | 2,66 | | 1,80 |

Wir können uns unter X, Y und Z die Verteilungen von fiktiven Stundenlöhnen für Gruppen aus unterschiedlichen Populationen vorstellen. Die Y-Verteilung unterscheidet sich von der X-Verteilung nur dadurch, daß die 4 Personen, die bisher 9 Geldeinheiten pro Stunde verdient haben, nun 10 Einheiten erhalten; dafür bekommen die 4 Personen mit dem höchsten Einkommen eine Geldeinheit weniger als zuvor. Die Umverteilung findet also nur in der oberen Hälfte der Einkommensverteilung statt. Arithmetisches Mittel und durchschnittliche Abweichung bleiben somit (bei gleicher Verteilungsmasse) unverändert. Die Streuung ist aber dennoch geringer geworden, ein Tatbestand, der durch den reduzierten s-Wert angezeigt wird, nicht aber durch AD.

In der Z-Verteilung sind sowohl die unteren als auch die oberen Einkommensklassen von X nicht mehr besetzt; die Umverteilung spielt sich in beiden Hälften der Einkommensverteilung ab. Die reduzierte Streuung spiegelt sich nun sowohl in s als auch in AD.

Dieser Vergleich zweier Kennzahlen soll über das gegebene Beispiel hinaus eine wichtige Funktion der Formalisierung theoretischer Konzepte verdeutlichen: Ausgangspunkt für die Definition von mittlerer Abweichung und Standardabweichung war ein zunächst nur diffuses Verständnis von »Streuung«. Der Versuch, diese Vorstellung zu formalisieren, führte zu zwei Alternativen, die, auf den ersten Blick besehen, vielleicht gleich-

wertig erschienen. Eine Analyse der formalen Eigenschaften zeigte jedoch wichtige Unterschiede auf. Jetzt muß der Theoretiker wiederum entscheiden, welche Formalisierung am ehesten seiner Vorstellung entspricht. Auf diese Weise wird er dazu gezwungen, seine Überlegungen in einer Richtung zu präzisieren, an die er zuvor vielleicht gar nicht gedacht hat. Wir können diesen Punkt weiter verdeutlichen, indem wir noch ein anderes Streuungsmaß einführen, den

**Variationskoeffizienten:**

Er sei hier mit V abgekürzt. Er setzt die Standardabweichung ins Verhältnis zum arithmetischen Mittel der Verteilung:

$$(3\text{-}9) \qquad V = \frac{s}{\bar{x}}$$

Die Besonderheit dieses Koeffizienten im Vergleich zur Standardabweichung läßt sich wie folgt verdeutlichen:

$$(3\text{-}10) \qquad \text{Aus } x' = x + c \quad \text{folgt} \quad s_{x'} = s_x$$

$$V_{x'} < V_x \text{ , falls } c > 0 \text{ und}$$

$$V_{x'} > V_x \text{ , falls } c < 0$$

Wenn zu allen Werten einer Verteilung eine Konstante c addiert wird, verändert sich der Mittelwert um diese Konstante, die Standardabweichung bleibt gleich. In V aber verändert sich der Nenner, folglich V selbst. Anders verhält es sich, wenn alle Werte mit einem bestimmten Faktor b multipliziert werden:

$$(3\text{-}11) \qquad \text{Aus } x' = bx \text{ folgt } s_{x'} = b \cdot s_x \text{ und}$$

$$(s_{x'})^2 = b^2 s_x^2, \text{ aber } V_{x'} = V_x$$

Wenn alle Werte sich um das b-Fache vergrößern, vervielfachen sich die **Differenzen** zwischen den einzelnen Werten um den Faktor b, die Streuung (Standardabweichung) wird in diesem Sinne größer. Die **Verhältnisse** zwischen den Werten verändern sich aber nicht. Wenn eine Person A doppelt soviel verdient hat wie Person B (z. B. 20 DM pro Stunde versus 10 DM pro Stunde), so gilt das Verhältnis von 2:1 auch nach der Multiplikation mit, beispielsweise, dem Faktor 1,5 [(30/15) = 2]. Wenn einem die Konstanz der Verhältnisse wichtiger ist als die Konstanz der absoluten Beträge, wird man beim Vergleich zweier Verteilungen den Variationskoeffizienten statt der Standardabweichung als Maßzahl nehmen. **Ob** einem die Konstanz der Proportionen wichtiger ist als die Konstanz der Differenzen, muß aber der Theoretiker (oder Politiker) entscheiden, nicht der

Statistiker. Der sagt nur, durch welche Maßzahl welches theoretische Konzept eher realisiert wird.

### 3.3 Momente (*)

Arithmetisches Mittel und Varianz lassen sich in die begrifflich übergeordnete Systematik der »Momente« einfügen. Wir wollen hier lediglich die Terminologie erläutern, ohne die Sache selbst näher zu erörtern. Man unterscheidet »Momente um den Ursprung« und »Momente um den Mittelwert« bzw. »zentrale Momente«. Das arithmetische Mittel wird als »Moment 1. Ordnung um den Ursprung« bezeichnet, die Varianz als »zentrales Moment 2. Ordnung«. Das »zentrale Moment 1. Ordnung« ist gleich Null definiert. Mit dieser Systematik gewinnt man zusätzliche Kenngrößen, die noch andere Aspekte der Verteilung summarisch beschreiben. Als zentrales Moment 3. Ordnung ist die Schiefe $m_3 = \Sigma\, z_i^3 / n$, $z_i = (x_i - x)/s$, definiert; das zentrale Moment 4. Ordnung, $m_4 = \Sigma\, z_i^4 / n$ liefert ein Maß für die Schmal- oder Breitgipfligkeit, den Grad der "Wölbung" einer Verteilung (auch "Exzeß" oder "Kurtosis").

# Kapitel 4

# Bivariate Verteilungen I: Elementare Tabellenanalyse und Korrelationskoeffizienten

Sozialwissenschaftliche Fragestellungen sind selten allein durch die Analyse univariater Verteilungen zu bearbeiten. Sie richten sich vielmehr auf vermutete **Beziehungen** zwischen zwei und mehr Variablen. Man möchte z. B. nicht nur wissen, wie sich die Frankfurter Nationalversammlung konfessionell zusammensetzte, sondern auch, ob das Abstimmungsverhalten der Abgeordneten mit ihrer Konfessionszugehörigkeit »zusammenhängt«. Des weiteren möchte man vielleicht untersuchen, ob der eventuelle Einfluß der Konfessionszugehörigkeit auf das Abstimmungsverhalten je nach Wahlregionen unterschiedlich war (siehe Kap. 5). Solche Fragen können nur mit Hilfe »gemeinsamer« Verteilungen (»joint distributions«) mehrerer Variablen beantwortet werden. Wir wollen in diesem und dem folgenden Kapitel einige elementare Techniken erläutern, derer man sich dabei bedienen kann.

In Kap. 4 beschränken wir uns auf die Betrachtung gemeinsamer Verteilungen von zwei Variablen (bivariate Verteilungen). Trivariate Verteilungen werden ausführlich in Kap. 5 erörtert.

In Abschn. 4.1 wird gezeigt, wie sich gemeinsame Verteilungen a) nichtmetrischer und b) metrischer Variablen anschaulich darstellen lassen. Die folgenden Abschnitte erläutern (zunächst nur deskriptiv benutzte) statistische Kennzahlen, die charakteristische Merkmale gemeinsamer Verteilungen zweier Variablen ausdrücken.

## 4.1 Darstellungsformen bivariater Verteilungen: Zweidimensionale Tabellen und Streudiagramme

Gemeinsame Verteilungen von Variablen mit nicht-metrischem Skalenniveau können in Form sog. **Kontingenztabellen** (Kreuz-, Assoziations-, Korrelationstabellen) dargestellt werden (Abschn. 4.1.1). Metrische Daten lassen sich durch Punkte in einem Koordinatenkreuz repräsentieren (Abschn. 4.1.2).

### 4.1.1 Zweidimensionale Tabellen: Struktur und Terminologie

Mit dem folgenden Schema (*Abb. 4.1*) und dem konkreten Beispiel (*Abb. 4.2*) läßt sich das Format zweidimensionaler Tabellen erläutern.

Die Tabelle besteht aus einer Kreuzung von r Zeilen ($r \geq 2$) und c Spalten ($c \geq 2$). Deshalb spricht man auch von einer (r x c)-Tabelle (row = Zeile,

Abb. 4.1: Schema des Aufbaus einer zweidimensionalen Tabelle

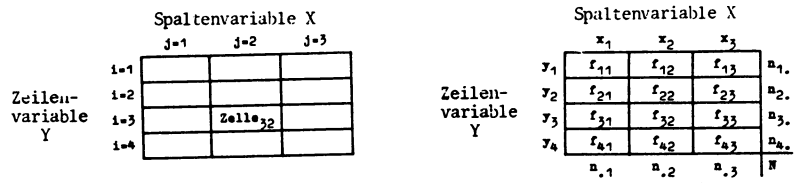



Abb. 4.2: Formale Schulbildung und Wanderungsintensität (Reichstagsabgeordnete 1912)

Schulbildung

| Wanderungs-intensitaet | COUNT | NIEDRIG 1 | MITTEL 2 | HOCH 3 | ROW TOTAL |
|---|---|---|---|---|---|
| NIEDRIG | 1 | 35 | 26 | 103 | 164 |
| | | | | | 39.3 |
| MITTEL | 2 | 38 | 15 | 113 | 166 |
| | | | | | 39.8 |
| HOCH | 3 | 15 | 11 | 61 | 87 |
| | | | | | 20.9 |
| COLUMN TOTAL | | 88 | 52 | 277 | 417 |
| | | 21.1 | 12.5 | 66.4 | 100.0 |

NUMBER OF MISSING OBSERVATIONS = 45

column = Spalte). Die Zahl der Spalten und Zeilen richtet sich nach der Zahl der Ausprägungen (evtl. nach Zusammenfassung vorher getrennter Kategorien) der beteiligten Variablen. In unserem Beispiel haben die Spaltenvariable »Schulbildung« (einschließlich Hochschulstudium) und die Zeilenvariable »Wanderungsintensität« jeweils drei Ausprägungen, nachdem einige Kategorien zusammengefaßt worden sind. Das ergibt eine

(3 x 3)-Tabelle mit neun »Zellen« (den inneren Feldern). Jede Zelle ist durch einen doppelten Index gekennzeichnet. Üblicherweise bezeichnet der erste Index »i« die durchnumerierten Ausprägungen der Zeilenvariable, der zweite Index »j« die Ausprägungen der Spaltenvariable. (Gelegentlich benutzen wir »i« auch als Fallindex und andere Buchstaben als Variablenindex; das ist aber aus dem jeweiligen Kontext ersichtlich.) Die Zahl $f_{32} = 11$ besagt also, daß 11 der klassifizierbaren Abgeordneten eine mittlere Schulbildung genossen haben **und** einen hohen Grad an Mobilität aufweisen (Aufenthalt im Ausland). Die an den Zeilen und Spaltenenden eingetragenen Häufigkeiten $n_{i.}$ und $n_{.j}$ bezeichnet man als **Randverteilungen** (»marginal distributions«). Das sind (im Falle zweidimensionaler Tabellen) die univariaten Häufigkeitsverteilungen der Zeilenvariable und der Spaltenvariable. Die Zusammenstellung der Häufigkeiten in einer Spalte oder einer Zeile nennt man **bedingte Häufigkeitsverteilungen** (»conditional distributions«). So finden wir in Spalte 2 unserer Beispieltabelle (*Abb. 4.2*) die Häufigkeitsverteilung der Wanderungsintensität ausschließlich für diejenigen Abgeordneten, die eine mittlere Schulbildung aufweisen. Mit anderen Worten: es handelt sich um die Häufigkeitsverteilung $f(y \mid x_2)$ der Zeilenvariable Y unter der Bedingung, daß in der Spaltenvariable X die zweite Ausprägung »realisiert« worden ist.

Welche Variable als Zeilen- und welche als Spaltenvariable eingesetzt wird, ist im Prinzip gleichgültig. Für den Fall, daß der Forscher eine Merkmalsdimension als die (kausal) bedingende (»unabhängige«) Variable X ansehen will, die einen »Einfluß« auf die andere Merkmalsdimension (»abhängige« Variable Y) ausübt, hat sich die Konvention eingebürgert, X als Spalten- und Y als Zeilenvariable einzusetzen. Wenn die Variablen auf nominalem Niveau gemessen worden sind, ist die Anordnung der jeweiligen Merkmalsausprägungen (der qualitativen Kategorien) gleichgültig. Bei ordinalen Daten muß die Rangordnung der Ausprägungen eingehalten werden. Dabei folgt man der Regel, die Rangstufen, beginnend mit dem niedrigsten Rang, von links nach rechts (Spaltenvariable) und von oben nach unten (Zeilenvariable) anzuordnen. Das ermöglicht es, Aussagen über »Beziehungen«, »Zusammenhänge« zwischen Variablen auf ein einheitliches Grundmuster zu beziehen (siehe vor allem Abschn. 4.2.3).

Die Tabelle in *Abb. 4.2* wurde mit dem Kommando

CROSSTABS TABLES = WANDERN BY SCHULB

erzeugt. Man kann mit dem Subkommando OPTIONS u. a. verschiedene Prozentuierungen (z. B. Spalten- oder Zeilenprozentuierung) und mit dem Subkommando STATISTICS verschiedene Assoziationsmaße anfordern. Davon werden wir später noch Gebrauch machen.

Auf den ersten Blick ist es nicht leicht, in unserer Beispieltabelle eine Beziehung zwischen den beteiligten Variablen zu entdecken oder eine entsprechende Hypothese als nicht bestätigt zu erkennen. Nimmt die Migrationsneigung mit dem Grad der Schulbildung zu oder nicht? Bei großen Tabellen (mit mehr Zellen) wird es im allgemeinen noch schwieriger, die Vielzahl der Informationen, die den Randverteilungen **und** den Zellenbesetzungen zu entnehmen sind, so zu verdichten, daß die für die jeweilige Fragestellung entscheidenden Merkmale deutlich hervortreten. Wie schon im univariaten Falle, helfen uns auch hier wieder spezifische Kennzahlen, von denen wir einige in Abschn. 4.2 besprechen werden.

### 4.1.2 Streudiagramme (»Scatterplots«)

Metrische Daten lassen sich nur dann in Tabellen darstellen, wenn sie gruppiert sind, also in wenigen Merkmalsklassen zusammengefaßt sind. Will man einen solchen Informationsverlust nicht als Voraussetzung der Darstellungsweise akzeptieren, bleibt der Ausweg des Streudiagramms (»Scatterplot«). *Abb. 4.3* zeigt die bivariate Verteilung der Variablen Y: Prozentanteile der SPD-Stimmen in den Wahlkreisen bei der Reichstagswahl 1912 und X: Prozentanteile der in Industrie und Gewerbe Beschäftigten (nebst Angehörigen) an der Bevölkerung im jeweiligen Wahlkreis.

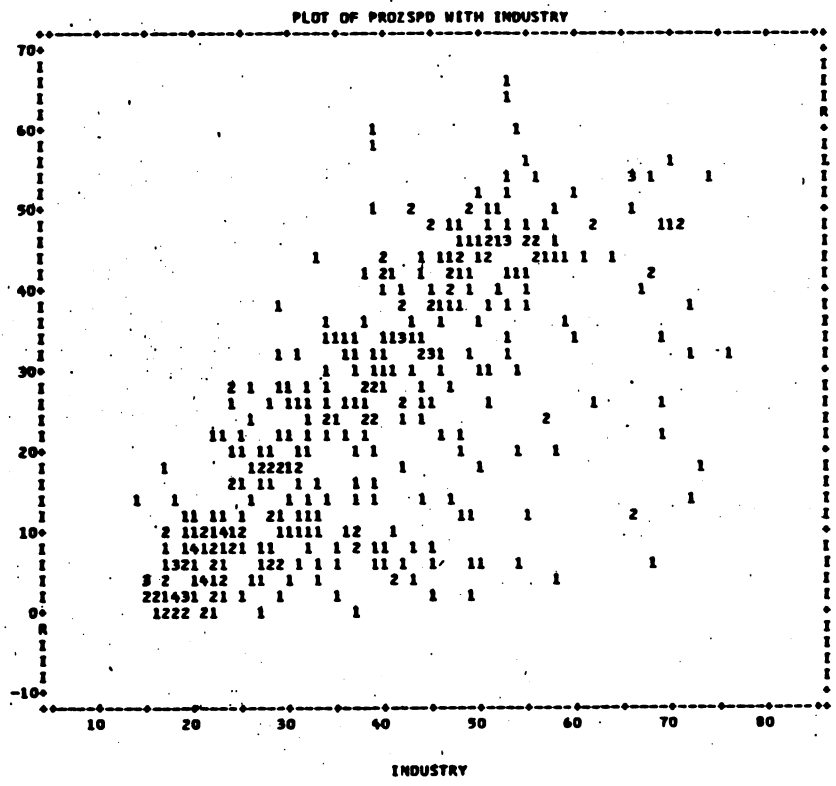
Ein solches Streudiagramm kann in SPSS$^x$ sowohl mit dem PLOT- als auch mit dem SCATTERGRAM-Befehl angefordert werden, in der einfachsten Version mit:

```
PLOT     /PLOT PROZSPD WITH INDUSTRY
```

Für jeden der n Wahlkreise ist a) auf der Abszisse der Wert $x_i$ des Industrialisierungsgrades, b) auf der Ordinate der Stimmenanteil $y_i$ des SPD-Kandidaten eingetragen ($i = 1,2,...,n$). Die Prozentuierung bezieht sich auf die Zahl der Wahlberechtigten im Wahlkreis. Die von den Werten $x_i$ und $y_i$ ausgehenden Koordinaten bilden Schnittpunkte $(x_i;y_i)$, die jeweils einen Wahlkreis repräsentieren. Im Plot geben die Ziffern an, wieviele Fälle (ungefähr) die gleiche Position in dem »Punktschwarm« einnehmen. Wir folgen der Konvention, die »unabhänige« Variable durch die Abszisse, die »abhängige« Variable durch die Ordinate darzustellen. Im Unterschied zu den vorangegangenen Beispielen sind die Untersuchungseinheiten diesmal nicht Personen (Individualdaten), sondern Wahlkreise (Kollektive, Aggregatdaten).

Das abgebildete Streudiagramm zeigt, wenig überraschend, daß die SPD um so mehr Stimmen erhielt, je größer der Anteil der in Industrie und Gewerbe Beschäftigten war. In Abschn. 4.2.4 und in Kap. 10 werden wir Verfahren erläutern, einen solchen Zusammenhang durch statistische Kennzahlen zu charakterisieren.

Abb. 4.3: Prozentanteile der SPD (Ordinate) und Industrialisierungsgrad der Wahlkreise (Abszisse)

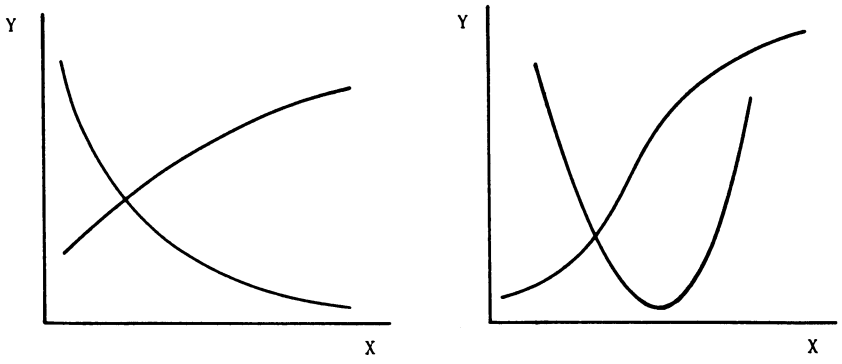



## 4.2 Statistische Kennziffern für den »Zusammenhang« zweier Variablen

Mit bi- und mehrvariaten Verteilungen will man die »Struktur« der »Beziehung« zwischen mehreren Variablen aufdecken. Nun kann man aber eine »Beziehung« in recht unterschiedlicher Weise theoretisch deuten und formulieren. So drückt z. B. die These: Je größer X, desto größer Y einen **linearen** Zusammenhang aus, in diesem Falle einen »positiven«. Die

These: Je größer X, desto kleiner Y zielt dagegen auf einen »negativen« linearen Zusammenhang. Für manche Variablenbeziehungen wird man aber eher einen nicht-linearen, »kurvenförmigen« Zusammenhang annehmen wollen, z. B. nach dem Muster von *Abb. 4.4*.

Abb. 4.4: Formen nicht-linearer Beziehungen




Gleichgültig, ob die Beziehung eine lineare oder irgendeine nicht-lineare Form hat, man wird in den Sozialwissenschaften nicht davon ausgehen, daß der Zusammenhang »deterministischer« Natur ist. Man wird »Ausnahmen«, »Fehler« einräumen, die These als »Tendenz« interpretieren, die Abweichungen zuläßt. Dann ist es aber auch nötig, von der unterschiedlichen »Stärke« eines Zusammenhangs zu sprechen und Maßzahlen zu entwickeln, die die Stärke relativ zur (linearen oder nicht-linearen) Form der Beziehung ausdrücken. Aussagen über lineare bzw. monotone Beziehungen setzen aber metrisches bzw. ordinales Meßniveau voraus. Was kann man also überhaupt unter einem »Zusammenhang« nominal skalierter Variablen verstehen? (siehe Abschn. 4.2.2)

Für jedes Meßniveau ist eine Vielzahl statistischer Maßzahlen entwikkelt worden, die jeweils unterschiedliche Aspekte gemeinsamer Verteilungen hervorheben. Das bedeutet, daß der Theoretiker sich über seine Hypothesen und Fragestellungen im klaren sein muß, wenn er sich für bestimmte Kennzahlen entscheidet. Es bedeutet aber auch, daß er über eini-

ge formale Eigenschaften der Maßzahlen Bescheid wissen muß, um sie theorieadäquat einsetzen zu können. Vielleicht wird er sogar erst über die Kenntnis formaler Eigenschaften statistischer Maßzahlen dazu angeregt, seine theoretischen Überlegungen zu präzisieren.

Wir beschränken uns in diesem Skript auf eine kleine Auswahl konventioneller Maßzahlen, die innerhalb der Sozialwissenschaften bis heute am häufigsten verwendet werden. (Einen breiter angelegten Überblick gibt z. B. Reynolds 1977a.) Komplexere Analysemodelle (siehe z. B. Hildebrand/Laing/Rosenthal 1977; 1977a), die teilweise auch die hier vorgestellten Maßzahlen mit enthalten (und unter neuen Gesichtspunkten deuten), teilweise auch alternative Koeffizienten definieren, sollen nicht dargestellt werden, weil das nur mit einem größeren technischen und mathematischen Aufwand möglich wäre. Das gilt erst recht für Modelle, die den regressionsanalytischen Ansatz (siehe Teil II) für die Analyse von Kontingenztabellen adaptieren (z. B. log-lineare Modelle).

### 4.2.1 Zum Einstieg: Die Prozentsatzdifferenz

Sicherlich ist die Prozentsatzdifferenz für den Nicht-Statistiker die anschaulichste und formal einfachste Größe, mit der ein Zusammenhang zweier Variablen ausgedrückt werden kann. Auch für den professionellen Datenanalytiker ist sie in vielen Fällen eine informative Maßzahl.

Beginnen wir mit dem Beispiel einer sog. Vier-Felder-Tafel, einer (2 x 2)-Tabelle, aus unserem Datensatz. Dazu haben wir die Variable Y: Konfessionszugehörigkeit der Reichstagsabgeordneten von 1912 dichotomisiert, indem wir die Protestanten allen anderen religiösen Gruppierungen gegenüberstellen. Die so rekodierte Variable wurde mit einer Standesvariablen X kreuztabelliert, die ebenfalls nur zwei Ausprägungen aufweist: »bürgerlich« oder »adlig« (s. *Abb. 4.5*).

Mit dieser bivariaten Verteilung soll die Frage beantwortet werden: Besteht ein Zusammenhang zwischen den beiden Variablen X und Y? Den Begriff des Zusammenhangs wird man hier so interpretieren, daß man fragt: Kommen bei den adligen Abgeordneten Protestanten häufiger oder weniger häufig vor als bei den bürgerlichen? Wir fragen also nach Unterschieden in den relativen Häufigkeiten, genauer: nach Unterschieden in der Verteilung der Y-Variable bei a) den Bürgerlichen, b) den Adligen. Dies ist die allgmeinste Interpretation des statistischen Zusammenhangsbegriffs: Ein Zusammenhang zwischen zwei Variablen besteht dann, wenn ihre bedingten Verteilungen der relativen Häufigkeiten unterschiedlich sind, wenn also Subgruppendifferenzen auftreten. Oder anders ausgedrückt: Falls kein Zusammenhang zwischen zwei Variablen besteht, sind die bedingten Verteilungen gleich und mit den jeweiligen Randverteilun-

Abb. 4.5: Bivariate Verteilung von Konfession und "Stand" der Reichstagsabgeordneten von 1912 (Spaltenprozentuierung)

| | | Stand | | |
|---|---|---|---|---|
| | COUNT / COL PCT | BUERGERL IICH | ADELIG | ROW TOTAL |
| | | 0 | 1. | |
| Konfession | | | | |
| 1. PROTESTANTISCH | | 172 | 40 | 212 |
| | | 43.4 | 60.6 | 45.9 |
| 2. ANDERE | | 224 | 26 | 250 |
| | | 56.6 | 39.4 | 54.1 |
| COLUMN TOTAL | | 396 | 66 | 462 |
| | | 85.7 | 14.3 | 100.0 |

gen identisch. Während man aber die »perfekte« Unabhängigkeit zweier Variablen über die Identität der bedingten Verteilungen exakt definieren kann, läßt sich eine »vollständige« Abhängigkeit nur für Variablen exakt definieren, die die gleiche Zahl von Ausprägungen aufweisen, in ihrer Kreuzklassifikation also quadratische Tabellen bilden. Mit »bedingt« ist hier nicht ein kausaler Einfluß gemeint, sondern nur eine verteilungsstatistische Beobachtung gemeint: Es handelt sich um die Häufigkeitsverteilung der einen Variablen Y, wie sie bei denjenigen Untersuchungsobjekten beobachtet wird, die alle bei der anderen Variablen X dieselbe Ausprägung $x_j$ aufweisen.

Dieser allgemeinsten Definition des Zusammenhangsbegriffs fügen die verschiedenen »Assoziations-« oder »Korrelationskoeffizienten« bestimmte mehr oder weniger einschneidende Spezifikationen (oder Restriktionen) hinzu. Aber Unterschiede der bedingten Verteilungen bleiben ein zentraler Bezugspunkt jeder Zusammenhangsanalyse. Die Frage ist also, wie man Unterschiede der bedingten Verteilungen zusammenfaßt und quantitativ präzisiert.

Im Falle einer Vier-Felder-Tafel ist das mit Hilfe der Prozentsatzdifferenz einfach zu bewerkstelligen. Relative Häufigkeiten lassen sich unmittelbar in prozentuierte Häufigkeiten umrechnen. Da jede der beiden Variablen nur mit zwei Ausprägungen vorkommt, kann der Unterschied der beiden bedingten Verteilungen mit einer einzigen Prozentsatzdifferenz ausgedrückt werden, denn die relativen Häufigkeiten müssen sich ja zu 1

bzw. zu 100 % ergänzen. In unserem Beispiel sind 60,6 % der adligen Abgeordneten, aber nur 43,4 % der bürgerlichen Abgeordneten Protestanten. Daraus ergibt sich eine Prozentsatzdifferenz von

(4-1) $d\% = 60{,}6 - 43{,}4 = 17{,}2$

Die zweite Prozensatzdifferenz: 39,4 - 56,6 = - 17,2 hat den gleichen Absolutbetrag, drückt also dieselbe Zusammenhangsstärke aus. Als Kennzahl hat d% klar definierte Obergrenzen, wie *Abb. 4.6* zeigt.

Derartige Verteilungen kommen natürlich in der Praxis des Sozialforschers kaum vor. Sie verdeutlichen aber die Grenzfälle des »perfekten« Zusammenhangs und des perfekten Nicht-Zusammenhangs. Ein perfekter Zusammenhang kann nur vorliegen, wenn die Randverteilungen der beiden Variablen gleich sind. Das Vorzeichen gibt lediglich an, ob die Fälle sich entlang der sog. Hauptdiagonalen von links oben nach rechts unten (siehe Tab. a) in *Abb. 4.6*) oder entlang der Nebendiagonalen von links unten nach rechts oben (siehe Tab. b) in *Abb. 4.6*) konzentrieren. Da die Kategorien dichotomer oder nominaler Variablen jederzeit umzustellen sind, kann das Vorzeichen nur in bezug auf eine gegebene Kategorienanordnung interpretiert werden.

Die Prozentsatzdifferenz ist eine »asymmetrische« Maßzahl. Wenn die beiden Randverteilungen nicht gleich sind, ist sie davon abhängig, welche der beiden Variablen als Basis für das Prozentuieren benutzt wird. In unserem Falle haben wir »in Richtung« der Spaltenvariable prozentuiert. Wenn wir in Richtung der Zeilenvariable prozentuieren, erhalten wir die Tabelle in *Abb. 4.7*. Die Prozentsatzdifferenz beträgt nun d% = 8,5.

Falls man die Beziehung kausal interpretieren, eine Variable als unabhängige, die andere als abhängige ansehen möchte, muß man die Randhäufigkeiten der unabhängigen Variablen X als Prozentuierungsbasis verwenden, unabhängig davon, ob sie als Spalten- oder als Zeilenvariable fungiert. Denn die Hypothese, daß X auf Y »wirkt«, können wir nur überprüfen, indem wir die bedingten relativen Y-Häufigkeiten in den einzelen Kategorien von X miteinander vergleichen. Keineswegs kann man d% oder andere asymmetrische Kennzahlen als Indikatoren für die **Richtung** einer eventuell vorliegenden Kausalbeziehung heranziehen. Welcher der beiden Koeffizienten (bei Vertauschen von abhängiger und unabhängiger Variabler) größer oder kleiner ist, hängt nur von den jeweiligen Randverteilungen ab, nicht von der Richtung des kausalen Einflusses[2].

[2] Es sei schon an dieser Stelle auf ein inferenzstatistisches Problem hingewiesen: Wenn eine Maßzahl wie d% auf ihre statistische »Signifikanz« überprüft werden soll (siehe Teil II, Kap. 8), sucht man allzu schiefe Verteilungen der Randhäufigkeiten zu vermeiden (evtl. indem man schwach besetzte Kategorien zusammenlegt). Als Daumenregel gilt, daß die einzelnen Kategorien nicht ungleicher als im Verhältnis 1:10 besetzt sein sollten.

Abb. 4.6: Extremfälle von d% (Spaltenprozentuierung)

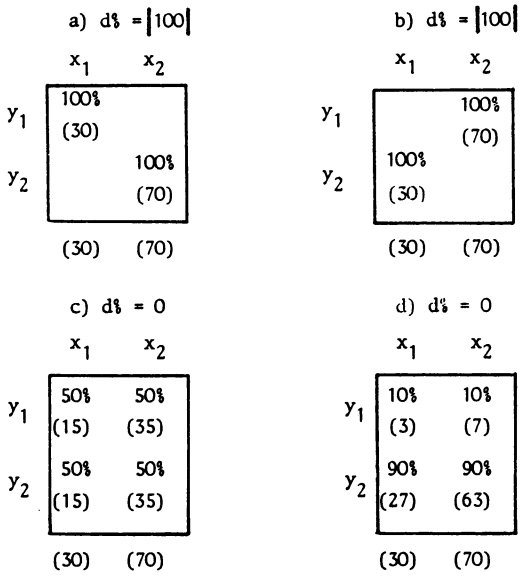



Abb. 4.7: Bivariate Verteilung von Konfession und "Stand" der Reichstagsabgeordneten von 1912 (Zeilenprozentuierung)

| | | Stand | | |
|---|---|---|---|---|
| | COUNT / ROW PCT | BUERGERLICH | ADELIG | ROW TOTAL |
| Konfession | 1. PROTESTANTISCH | 172 / 81.1 | 40 / 18.9 | 212 / 45.9 |
| | 2. ANDERE | 224 / 89.6 | 26 / 10.4 | 250 / 54.1 |
| | COLUMN TOTAL | 396 / 85.7 | 66 / 14.3 | 462 / 100.0 |

Wenn die Tabelle mehr als vier Felder umfaßt, lassen sich die Unterschiede zwischen den bedingten Verteilungen nicht mehr in einer einzigen Prozentsatzdifferenz ausdrücken. Die folgende Tabelle in *Abb. 4.8* enthält die Kreuzklassifikation der Adelsvariable mit einer Konfessionsvariable, die fünf Ausprägungen enthält.

Abb. 4.8: Bivariate Verteilung von Konfession und "Stand" der Reichstagsabgeordneten von 1912 (Spalten- und Zeilenprozentuierung)

| | | Stand | | |
|---|---|---|---|---|
| | COUNT | I | | |
| | ROW PCT | IBUERGERL | ADELIG | ROW |
| | COL PCT | IICH | | TOTAL |
| | | I 0I | 1.I | |
| Konfession | | I | I | |
| 1. PROTESTANTISCH | | I 173 | I 40 | I 213 |
| | | I 81.2 | I 18.8 | I 46.1 |
| | | I 43.7 | I 60.6 | I |
| 2. KATHOLISCH | | I 128 | I 26 | I 154 |
| | | I 83.1 | I 16.9 | I 33.3 |
| | | I 32.3 | I 39.4 | I |
| 3. MOSAISCH | | I 9 | I 0 | I 9 |
| | | I 100.0 | I 0 | I 1.9 |
| | | I 2.3 | I 0 | I |
| 6. DISSIDENT | | I 76 | I 0 | I 76 |
| | | I 100.0 | I 0 | I 16.5 |
| | | I 19.2 | I 0 | I |
| 8. ANDERE | | I 10 | I 0 | I 10 |
| | | I 100.0 | I 0 | I 2.2 |
| | | I 2.5 | I 0 | I |
| | COLUMN | 396 | 66 | 462 |
| | TOTAL | 85.7 | 14.3 | 100.0 |

Hier können wir bereits vier unterschiedliche d%-Maße bilden (mit den Häufigkeiten der Standesvariable als Prozentbasis). Hätte die Standesvariable drei Ausprägungen, müßten drei Paare bedingter Verteilungen mit Hilfe von dreimal vier Prozentsatzdifferenzen miteinander verglichen werden. Man könnte auf den Gedanken kommen, ein einziges summarisches Maß dadurch zu erhalten, daß man den Durchschnitt aller d%-Werte errechnet. Das wäre aber insofern unbefriedigend, als sich die Prozentsatzdifferenzen jeweils auf unterschiedliche Fallzahlen bezögen. Selbst wenn man Gewichtungsregeln fände, die die Ober- und Untergrenze bewahrten, bliebe offen, ob die neue Maßzahl auch in ein inferenzstatistisches Modell

integriert werden könnte (was anzustreben ist, wie in Teil II deutlich werden wird). Wir wollen diese Frage hier nicht weiter verfolgen. Die Statistiker haben für größere Tabellen andere Kennzahlen vorgeschlagen, übrigens auch andere (zusätzliche) Maßzahlen für die gemeinsame Verteilung zweier dichotomer Merkmale. Einige davon werden wir in den nächsten Abschnitten vorstellen.

### 4.2.2 Nominales Meßniveau: Zusammenhangsmaße auf der Basis von Chi-Quadrat

Statt bedingte relative Häufigkeitsverteilungen **untereinander** zu vergleichen, also Subgruppendifferenzen zu ermitteln, kann man die beobachteten Häufigkeiten auch mit einer **theoretisch** bestimmten Größe vergleichen. Man kann z. B. fragen: Mit welchen Häufigkeiten müßten die einzelnen Zellen der Tabelle besetzt sein, wenn überhaupt kein Zusammenhang zwischen den Variablen bestünde, wenn sie unabhängig voneinander wären? Kann man diese Frage beantworten, so lassen sich die unter der Unabhängigkeitsthese »erwarteten« Häufigkeiten mit den beobachteten Häufigkeiten vergleichen. Die dabei ermittelten Differenzenbeträge (die noch in einer bestimmten Weise zusammengefaßt und standardisiert werden müssen) erlauben Schlußfolgerungen über die Stärke eines Zusammenhangs.

Mit dem Konzept der statistischen Unabhängigkeit eilen wir einem Thema voraus, das ausführlich erst in den Kapiteln 6 und 7 (Teil II) erörtert wird. Andererseits können wir anhand eines konkreten Anwendungsbeispiels die formale Definition dieses Begriffs anschaulich vorbereiten. Als Beispiel betrachten wir die Kreuztabellierung der Standes- und der Konfessionsvariable, deren Ausprägungen wir in jeweils drei Kategorien zusammenfassen (*Abb. 4.9*).

Die Tabelle enthält die beobachteten absoluten Häufigkeiten sowie die prozentuierten Häufigkeiten. Die erste Prozentangabe in einer Zelle resultiert aus einer Prozentuierung zur Basis der Zeilenvariable, die zweite Angabe bezieht sich auf die Basis der Spaltenvariable. Unter den 443 erfaßten Abgeordneten sind zum Beispiel 213 Protestanten und 378 Bürgerliche. Wie groß müßte die Zahl derer sein, die sowohl protestantisch als auch bürgerlich sind, wenn zwischen den beiden Variablen kein Zusammenhang bestünde?

Nach unserer allgemeinen Definition des Zusammenhangs müßten alle bedingten Verteilungen der relativen Häufigkeiten gleich, also auch identisch mit der betreffenden Randverteilung sein, falls die beiden Variablen (völlig) unabhängig voneinander wären. Wir können also die relative bzw. prozentuierte Randverteilung der einen Variable, sagen wir der Konfes-

Abb. 4.9: Bivariate Verteilung von Konfession und "Stand" der Reichstagsabgeordneten von 1912

| | | Konfession | | | |
|---|---|---|---|---|---|
| | COUNT | 1 | 2 | 6 | |
| | ROW PCT | PROTESTA | KATHOL | ANDERE | ROW |
| | COL PCT | INTISCH | | T | TOTAL |
| Stand | | 1 | 2 | 6 | |
| 0 | | 174 | 128 | 76 | 378 |
| BUERGERLICH | | 46.0 | 33.9 | 20.1 | 85.3 |
| | | 81.7 | 83.1 | 100.0 | |
| 1 | | 24 | 10 | 0 | 34 |
| ADELIG | | 70.6 | 29.4 | 0 | 7.7 |
| | | 11.3 | 6.5 | 0 | |
| 2 | | 15 | 16 | 0 | 31 |
| HOCHADEL | | 48.4 | 51.6 | 0 | 7.0 |
| | | 7.0 | 10.4 | 0 | |
| COLUMN TOTAL | | 213 | 154 | 76 | 443 |
| | | 48.1 | 34.8 | 17.2 | 100.0 |

RAW CHI SQUARE = 19.88183 WITH
CONTINGENCY COEFFICIENT = .20725

NUMBER OF MISSING OBSERVATIONS = 19

sionsvariable, als Maßstab oder Kriterium für ihre bedingten Häufigkeitsverteilungen in jeder der drei Standesgruppen benutzen. Wenn kein Zusammenhang bestünde, müßten also bei den Bürgerlichen (ebenso wie bei denen aus dem niederen und dem hohen Adel) 48,1 % protestantisch, 34,8 % dagegen katholisch sein und 17,2 % keiner dieser beiden Konfessionen angehören. Im linken oberen Feld der Tabelle würden wir also

$$(4\text{-}2) \qquad f_e = \frac{378 \times 48{,}1}{100} = 181{,}8$$

Abgeordnete erwarten. Auf die gleiche Zahl kommen wir auch durch folgende Rechnung:

$$(4\text{-}3) \qquad f_e = \frac{378 \times 213}{443} = 181{,}7$$

Wir multiplizieren also die beiden Randhäufigkeiten und dividieren das Produkt durch die Gesamtzahl der Fälle, allgemein:

$$(4\text{-}4) \qquad f_e = \frac{n_{j.} \times n_{.i}}{N}$$

Auf diese Weise können wir für jede Zelle die unter der Unabhängigkeitsthese erwartete Häufigkeit berechnen.

Die Rechenformel (4-4) läßt sich wahrscheinlichkeitstheoretisch interpretieren: Wenn unter 443 Abgeordneten 378 Bürgerliche sind, dann gäbe es eine Wahrscheinlichkeit P(B) = 378/443, daß wir bei einer rein zufallsgesteuerten Auswahl aus der Gesamtheit der Abgeordneten einen aus dem bürgerlichen Lager »treffen« würden. Die Wahrscheinlichkeit, einen Protestanten zufällig »herauszuziehen«, wäre P(Pr) = 213/443. Die Wahrscheinlichkeit, zufällig einen Abgeordneten auszuwählen, der sowohl bürgerlich als auch protestantisch ist, wäre nach dem sog. Multiplikationstheorem der Wahrscheinlichkeitsrechnung (siehe Kap. 6, Teil II)

$$(4\text{-}5) \qquad P(B + Pr) = (378/443) \cdot (213/443)$$

Wenn wir diesen Zufallsvorgang 443mal unabhängig voneinander wiederholen könnten, wäre damit zu rechnen, daß wir

$$(4\text{-}6) \qquad \frac{378 \times 213 \times 443}{443 \times 443} = \frac{378 \times 213}{443} = 181{,}7 \approx 182$$

protestantisch-bürgerliche Abgeordnete erhielten. Diese Gleichung (4-6) ist identisch mit (4-3).

Wenn wir die erwarteten Häufigkeiten $f_e$ für alle Zellen zusammenstellen, erhalten wir die sog. **Indifferenztabelle** (siehe *Abb. 4.10*).

Sie ist mit der Tabelle zu vergleichen, die die beobachteten Häufigkeiten $f_b$ enthält und die man in diesem Zusammenhang als **Kontingenztabelle** bezeichnet (siehe *Abb. 4.9*)

Der Vergleich der beiden Tabellen geschieht numerisch nach folgender Formel:

$$(4\text{-}7) \qquad \sum_{k=1}^{K} \frac{\left(f_{b_k} - f_{e_k}\right)^2}{f_{e_k}} = \chi^2 = 19.88$$

Abb. 4.10: Indifferenztabelle zur Kontingenztabelle in Abb. 4.9

| | Prot. | Kath. | Andere | |
|---|---|---|---|---|
| Bürgerl. | 181,7 | 131,4 | 64,8 | 377,9 |
| Adel | 16,3 | 11,8 | 5,8 | 33,9 |
| Hochadel | 14,9 | 10,8 | 5,3 | 31,0 |
| | 212,9 | 154,0 | 75,9 | 442,8 |

Abb. 4.11: Chi-Quadrat-Werte von Tabellen mit unterschiedlichen Fallzahlen und gleichbleibenden Prozentdifferenzen

| | | |
|---|---|---|
| 20 | 10 | 30 |
| 10 | 20 | 30 |
| 30 | 30 | 60 |

$\chi^2 = 6.66$

| | | |
|---|---|---|
| 40 | 20 | 60 |
| 20 | 40 | 60 |
| 60 | 60 | 120 |

$\chi^2 = 13.33$

| | | |
|---|---|---|
| 60 | 30 | 90 |
| 30 | 60 | 90 |
| 90 | 90 | 180 |

$\chi^2 = 19.99$

Man bildet für jede der K Zellen die Differenz zwischen erwarteter und beobachteter Häufigkeit und quadriert sie. Dadurch wird vermieden, daß sich positive und negative Beträge bei der Summenbildung ausgleichen. Dann »relativiert« oder »standardisiert« man das Gewicht dieser Differenz durch die jeweils erwartete Häufigkeit. Ein bestimmter Differenzbetrag zählt um so mehr, je kleiner die erwartete Häufigkeit ist.

Die Größe Chi-Quadrat ist aber noch kein geeignetes Maß für die Stärke des Zusammenhangs. Wie die Tabellen in *Abb. 4.11* zeigen, ist Chi-Quadrat abhängig von der Zahl N der Fälle.

Identische konditionale Verteilungen relativer Häufigkeiten führen bei einer Vervielfachung der Zellenhäufigkeiten zu einer Vervielfachung von Chi-Quadrat um den gleichen Faktor. Da Forscher ständig Tabellen vergleichen, die auf unterschiedlichen Fallzahlen beruhen, ist dies natürlich ein unerwünschter Effekt. Die Statistiker haben im Laufe der Zeit eine ganze Reihe von Vorschlägen gemacht, wie man die Größe $\chi^2$ weiter bearbeiten kann, um diesen Effekt zu beseitigen.

Der vielleicht naheliegendste Vorschlag ist, $\chi^2$ einfach durch die Zahl der Fälle zu dividieren. Die so konstruierte Maßzahl bezeichnet man als **Phi-Koeffizienten**:

$$(4\text{-}8) \qquad \varphi^2 = \frac{\chi^2}{N} \;,\; \varphi = \sqrt{\frac{\chi^2}{N}} \;,$$

In einer (2 x 2)-Tabelle erreicht $\varphi$ (bzw. $\varphi^2$) den maximalen Wert »1«, wenn zwei Diagonalzellen unbesetzt sind. Das ist nur möglich, wenn die Randverteilungen der Zeilen- und der Spaltenvariable gleich sind. Für größere als (2 x 2)-Tabellen kann Phi größer als 1 werden, wodurch es als Vergleichsgröße praktisch untauglich wird. Aber auch seine Anwendung bei der Analyse von 4-Felder-Tafeln hat ihre Tücken. Der Koeffizient ist nicht »stabil«, er reagiert empfindlich auf Veränderungen in den Randverteilungen. Wenn keine statistische Unabhängigkeit besteht, wird das maximale Phi um so kleiner, je unterschiedlicher die Randverteilungen der Spalten- und der Zeilenvariable sind. Das maximale Phi ist wie folgt bestimmt (siehe Guilford 1954, S. 358 f):

$$(4\text{-}8a) \qquad \varphi_{max} = \sqrt{\left(\frac{p_j}{q_j} \cdot \frac{q_i}{p_i}\right)} \quad , \; p_i \geq p_j \;.$$

Dabei bezeichnen $p_i$ und $p_j$ die jeweils größte relative Häufigkeit in den beiden Variablen der 4-Felder-Tabelle und $q = 1\text{-}p$. Nur wenn $p_i = p_j$, wird der Wurzelausdruck gleich 1. *Abb. 4.12* bringt einige fiktive Zahlenbeispiele.

Abb. 4.12: Abhängigkeit des Phi-Koeffizienten von den Randverteilungen

a) Gleiche Randverteilungen von X und Y

| | | |
|---|---|---|
| 50 | 0 | 50 |
| 0 | 50 | 50 |
| 50 | 50 | 100 |

$\varphi = 1$

| | | |
|---|---|---|
| 10 | 0 | 10 |
| 0 | 90 | 90 |
| 10 | 90 | 100 |

$\varphi = 1$

b) Ungleiche Randverteilungen von X und Y

| | | |
|---|---|---|
| 30 | 0 | 30 |
| 20 | 50 | 70 |
| 50 | 50 | 100 |

$$\varphi_{max} = \sqrt{\frac{0.5 \cdot 0.3}{0.5 \cdot 0.7}} = 0.655$$

| | | |
|---|---|---|
| 10 | 0 | 10 |
| 40 | 50 | 90 |
| 50 | 50 | 100 |

$$\varphi_{max} = \sqrt{\frac{0.5 \cdot 0.1}{0.5 \cdot 0.9}} = 0.33$$

Die Randverteilungen kann sich ein Sozialwissenschaftler selten aussuchen. Daß in dem Reichstag von 1912 viel weniger Adlige saßen als im Parlament von 1871 kann er nicht ändern. Bei Vergleichen von Zusammenhangsstärken in unterschiedlichen Tabellen sollte man also darauf achten, wie unterschiedlich die Randverteilungen sind und ob der verwendete Koeffizient abhängig ist von den Randverteilungen. Beispielsweise sind Maßzahlen, die auf Chi-Quadrat beruhen, alle abhängig von Randverteilungen. Das gilt allerdings in mehr oder weniger starker Ausprägung für fast alle Zusammenhangsmaße, die auf dem Markt sind. In den meisten Fällen kann man sich helfen, indem man Tabellen mit unterschiedlichen Randverteilungen »standardisiert«. Das geschieht z. B., indem man sämtliche Zellenhäufigkeiten zur Basis der Randhäufigkeiten der unabhängigen Variable prozentuiert (damit die Randhäufigkeiten praktisch alle gleich 100 setzt) und mit den Prozenthäufigkeiten, statt der absoluten Häufigkeiten weiterreichnet (siehe Reynolds 1977, 17 f., 50; Liebetrau 1983, 88). Die Prozentdifferenz reagiert nicht auf Veränderungen in

der Randverteilung der unabhängigen Variablen. Es sind auch iterative Verfahren vorgeschlagen worden, die eine Tabelle hinsichtlich der Randverteilungen **beider** Variablen standardisieren (siehe Reynolds 1977 a, S. 32 ff.). Ein Koeffizient für 4-Felder-Tafeln, der unabhängig ist von Randverteilungsänderungen, ist »Yules Q«, ein Spezialfall von »Gamma« (siehe Abschnitt 4.2.3.1).

Phi ist übrigens identisch mit dem in Abschn. 4.2.4 zu besprechenden Produkt-Moment-Korrelationskoeffizienten von Pearson, wenn die beiden Ausprägungen der dichotomen Variablen mit »0« und »1« kodiert sind.

Eine bei sozialwissenschaftlichen Praktikern (zu Unrecht) immer noch sehr beliebte Maßzahl für die Stärke eines Zusammenhangs ist der von Pearson entwickelte **Kontingenzkoeffizient C**:

$$(4\text{-}9) \qquad C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = 0{,}21$$

Er hat zwar eine klar definierte Untergrenze, 0, erreicht aber nie den Wert 1, nähert sich ihm nur mit größer werdender Tabelle an. Der maximale Wert, der sich nur für quadrierte Tabellen angeben läßt, ist:

$$(4\text{-}10) \qquad C_{max} = \sqrt{\frac{r-1}{r}} = \sqrt{\frac{c-1}{c}} = \sqrt{\frac{3-1}{3}} = 0{,}816$$

(für eine 3×3 Tabelle)

Für nicht-quadratische Tabellen gilt das zweite Gleichheitszeichen nicht. Man behilft sich, indem man zwei $C_{max}$-Werte ausrechnet und anschließend das arithmetische Mittel daraus bildet. Aber das sind ziemlich fragwürdige Behelfskrücken.

In der Fachliteratur eher akzeptiert ist eine weitere auf $\chi^2$ beruhende Maßzahl, die als **Cramers V** bezeichnet wird:

$$(4\text{-}11) \qquad V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1, c-1)}} = 0{,}15$$

Dieser Koeffizient hat den Vorteil, daß er für Tabellen unterschiedlicher Größe den Wert 1 annehmen (und nicht überschreiten) kann, sowohl bei quadratischen als auch bei rechteckigen (r # c) Tabellen. V ist also eine brauchbare Vergleichsgröße für die Stärke eines Zusammenhangs in unterschiedlichen Tabellen, solange sich die Randverteilungen nicht stark voneinander unterscheiden.

Angewandt auf 4-Felder-Tafeln ist V identisch mit Phi.

Wenn man auf der Basis von Chi-Quadrt statistische Tests durchführt (siehe Kap. 8), müssen hinsichtlich der Zellenbesetzungen bestimmte Voraussetzungen erfüllt sein, die wir in Abschnitt 8.7.2 (Teil II) erläutern.

Alle Zusammenhangsmaße, die auf Chi-Quadrat beruhen, kranken daran, daß ihre Werte zwischen Null und Eins inhaltlich nicht interpretierbar sind. Man kann lediglich sagen, der Zusammenhang (im Sinne der Abweichung von der statistischen Unabhängigkeit) sei um so stärker (kleiner), je größer (kleiner) der Koeffizient ist.

Eine inhaltlich interpretierbare Maßzahl für Nominaldaten haben Goodman und Kruskal mit dem Koeffizienten **Lambda** vorgeschlagen. Er hat aber ebenfalls Nachteile (beispielsweise kann er auch dann Null werden, wenn die bedingten Verteilungen nicht gleich sind), so daß er nur zusammen mit anderen Maßzahlen verwendet werden sollte. Ein Vorteil der Konstruktion von Indifferenztabellen liegt darin, daß sie den Vergleich erwarteter und beobachteter Häufigkeiten für **einzelne** Zellen ermöglichen. Man kann so leicht feststellen, in **welchen** Zellen große oder kleine Differenzen (relativ zum Erwartungswert) vorkommen; man muß (und sollte) sich nicht allein mit der summarischen Maßzahl zufriedengeben.

Die Möglichkeiten, interpretierbare Maßzahlen für Zusammenhänge zwischen nominalen Variablen zu konstruieren, sind dadurch begrenzt, daß sich eine (z. B. lineare oder monotone) »Form« des Zusammenhangs nicht formulieren läßt, da die Kategorien nicht in eine Größenordnung zueinander gebracht werden können, ihre Anordnung also beliebig ist. Dadurch fehlt der Aussage über die »Stärke« der Beziehung ein klarer inhaltlicher Bezugspunkt.

Klar definiert ist immerhin das (wahrscheinlichkeits-)theoretische Modell der statistischen Unabhängigkeit zweier Variablen, das zur Maßzahl $\chi^2$ führt. Wie groß die Differenz zwischen theoretischem Modell (den erwarteten Häufigkeiten) und empirischer Realität (den beobachteten Häufigkeiten) sein muß, um als »überzufällig« gelten zu können, werden wir in Teil II, Kap. 8 erörtern.

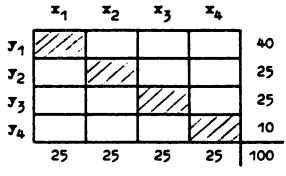
Mit diesem Beispiel ist eine typische Vorgehensweise der quantitativen Sozialforschung illustriert: Theoretische Hypothesen (wie die der Unabhängigkeit zwischen zwei Merkmalsdimensionen) werden in ein formales Modell »übersetzt«, das anschließend mit den empirischen Daten konfrontiert wird und sich dabei bewähren oder scheitern kann. In den folgenden Abschnitten werden den Modellen für Unabhängkeit oder Zusammenhang zweier Variablen weitere Elemente hinzugefügt. Sie erlauben es, die jeweilige numerische Ausprägung der Maßzahlen für die Zusammenhangsstärke inhaltlich zu interpretieren.

### 4.2.3 Proportionale Fehlerreduktion: Einige Zusammenhangsmaße für ordinale Variablen

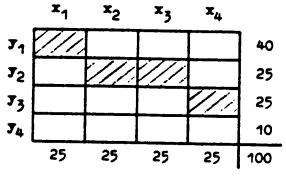
Liegen Rangskalen vor, läßt sich die Form einer Beziehung als monoton oder nicht-monoton kennzeichnen. Bei einer positiven monotonen Beziehung sind höhere Ränge in der X-Variablen tendenziell mit höheren oder gleichbleibenden Rängen der Y-Variablen verbunden. Bei negativen monotonen Beziehungen sind höhere Ränge in der X-Variablen tendenziell mit niedrigeren oder gleichbleibenden Rängen der Y-Variablen verknüpft. Nicht-monotone Beziehungen können U- oder V-förmig verlaufen. Schematisch lassen sich mögliche Beziehungsformen für Ordinalvariablen wie in *Abb. 4.13* kennzeichnen.

Abb. 4.13: Beziehungsmuster ordinaler Variablen

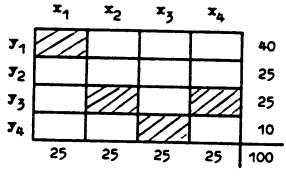
a) monotone Beziehung




b) monotone Beziehung




c) nicht-monotone Beziehung



Alle Maßzahlen für die Stärke eines Zusammenhangs, die wir in diesem Abschnitt besprechen, beziehen sich auf monotone Zusammenhänge. Sollte also ein nicht-monotoner Zusammenhang vorliegen, der theoretisch plausibel interpretierbar ist, so würde er durch diese Koeffizienten nicht erfaßt.

Wie lassen sich die Ranginformationen bei der Konstruktion von bivariaten Kennzahlen verwenden, obwohl die Rangdifferenzen nicht quantifizierbar sind? Zwei Konzeptionen sind für die hier vorgestellte Lösungsstrategie zentral: Erstens wird der Zusammenhangsbegriff im Sinne einer Prognosemöglichkeit interpretiert. Zweitens vergleicht man nicht Einzelfälle miteinander, sondern Paare. (Zur Kritik an dieser Konzeption siehe Wilson 1971; Hildebrand et al. 1977; 1977a.)

Wenn ein Zusammenhang zwischen zwei Variablen, X und Y, besteht, dann muß es möglich sein, die Werte der einen Variablen (Y) bei Kenntnis der Werte der anderen Variablen (X) besser zu prognostizieren, als das ohne diese Kenntnis möglich wäre. Gedanklich führen wir also zwei Prognosen durch:

(a) Die Prognose von Y-Werten ohne Kenntnis der X-Werte. Wir verhalten uns also wie Roulettespieler. Dabei unterlaufen uns bei wiederholten Versuchen Prognosefehler, deren Summe wir mit $E_1$ bezeichnen wollen.

(b) Die Prognose von Y-Werten auf der Basis der bekannten X-Werte. Wenn ein Zusammenhang zwischen X und Y besteht und wenn wir uns rational verhalten, dann müssen dabei weniger Fehler auftreten als zuvor: $E_2 < E_1$.

Wenn die beiden Fehlermengen bestimmbar sind, lassen sie sich in folgender Weise ins Verhältnis zueinander setzen:

$$\frac{E_1 - E_2}{E_1} = PRE \qquad (4\text{-}12)$$

Das heißt, es ließe sich die **proportionale Fehlerreduktion** («Proportional Reduction of Error«) feststellen, die man beim Übergang vom ersten zum zweiten Prognoseverfahren erreichen würde. Sie könnte als Maß für die Stärke des Zusammenhangs dienen, denn sie hätte klar definierte Unter- und Obergrenzen. Wenn $E_1 = E_2$, wenn sich also die Prognose der Y-Werte durch die Kenntnis der X-Werte nicht verbessern ließe, gäbe es offensichtlich keinen Zusammenhang der beiden Variablen und PRE wäre folgerichtig null. $E_2 = 0$ würde bedeuten, daß bei Kenntnis der X-Werte die Y-Werte fehlerfrei vorausgesagt werden könnten; es bestünde ein perfekter Zusammenhang. PRE würde in diesem Falle zu $E_1/E_1 = 1$.

Geklärt werden müssen nun vor allem die Regeln, die angeben, wie die Prognosen im einzelnen vorzunehmen sind und welche Fehlerdefinitionen daraus folgen. Da die Prognosen nur gedanklich durchgeführt werden, dienen diese Regeln nicht der praktisch-technischen Instruktion, sondern explizieren eine bestimmte Deutung des Zusammenhangsbegriffs, mit anderen Worten: sie bilden ein (theoretisches) Modell.

Eine (nicht konkurrenzlose aber meistangewandte) Strategie hat ihren Ausgangspunkt im Konzept der »Paare«, das wir als erstes erläutern wollen. Nehmen wir an, wir ziehen aus unserer Stichprobe (oder Grundgesamtheit) einen beliebigen Fall A heraus und registrieren seinen Rangplatz sowohl auf der X- als auch auf der Y-Variable: $(X_A;Y_A)$. Nun wählen wir einen zweiten beliebigen Fall, B, aus und registrieren ebenfalls dessen X- und Y-Ränge: $(X_B;Y_B)$. Die X- und Y-Werte dieses Paares (A,B) von Untersuchungseinheiten können unterschiedliche Relationen bilden. Wir wollen das anhand eines Beispiels aus unserem Datensatz demonstrieren.

Dazu präsentieren wir erneut die Tabelle, in der die Variablen X = Schulbildung und Y = Migrationsgrad kreuzklassifiziert und die Zellen fortlaufend mit Buchstaben identifiziert sind (*Abb. 4.14*).

Abb. 4.14: Formale Schulbildung und Wanderungsintensität (Reichstagsabgeordnete 1912)

| | | Schulbildung | | | |
|---|---|---|---|---|---|
| | COUNT | INIEDRIG 1 | MITTEL 3 | HOCH 5 | ROW TOTAL |
| Wanderungs-intensitaet | | | | | |
| 1 | | 35 | 26 | 103 | 164 |
| NIEDRIG | | a | b | c | 39.3 |
| 2 | | 38 | 15 | 113 | 166 |
| MITTEL | | d | e | f | 39.8 |
| 3 | | 15 | 11 | 61 | 87 |
| HOCH | | g | h | i | 20.9 |
| | COLUMN TOTAL | 88 21.1 | 52 12.5 | 277 66.4 | 417 100.0 |

NUMBER OF MISSING OBSERVATIONS = 45

Folgende Paarbeziehungen sind möglich:

(a) Die Untersuchungseinheiten können im Hinblick auf X und Y »gleichsinnig« geordnet sein, z. B.

(4-13) $(X_A < X_B)$ und $(Y_A < Y_B)$

Der zweite Fall (B) nimmt also nicht nur auf der X-, sondern auch auf der Y-Dimension den höheren Rangplatz ein. »Gleichsinnig« wäre auch folgende Anordnung:

(4-13') $(X_A > X_B)$ und $(Y_A > Y_B)$

Solche Relationen bezeichnet man als **konkordante Paare.** Ihre Gesamtzahl wird mit dem Symbol $N_c$ angegeben.
Ein konkordantes Paar bilden beispielsweise ein Abgeordneter aus Zelle a und ein Abgeordneter aus Zelle e.

(b) Die Untersuchungseinheiten können im Hinblick auf X und Y »gegensinnig« geordnet sein:

(4-14) $(X_A < X_B)$ und $(Y_A > Y_B)$ ; $(X_A > X_B)$ und $(Y_A < Y_B)$

Wer auf der X-Variablen den niedrigeren (höheren) Rangplatz einnimmt, hat auf der Y-Variablen den höheren (niedrigeren) Rang. Solche Relationen bezeichnet man als **diskordante Paare,** $N_d$. Ein solches Paar bilden beispielsweise ein Abgeordneter aus Zelle c und ein Abgeordneter aus Zelle e.

(c) Die Untersuchungseinheiten können in ihren X-Werten gleich (verknüpft, gebunden, »tied«), in ihren Y-Werten jedoch verschieden sein:

(4-15) $(X_A = X_B)$ und $(Y_A \neq Y_B)$

Die Menge dieser (einseitig) »in X verknüpften Paare« erhält das Symbol $T_x$ (»tied in x«). Ein solches Paar bilden beispielsweise ein Abgeordneter aus Zelle a und ein Abgeordneter aus Zelle d.

(d) Die Untersuchungseinheiten können sich in ihren X-Rangplätzen unterscheiden, aber den gleichen Y-Rangplatz einnehmen:

(4-15a) $(X_A \neq X_B)$ und $(Y_A = Y_B)$

Diese (einseitig) »in Y verknüpften Paare« erhalten das Symbol $T_y$. Ein solches Paar bilden beispielsweise ein Abgeordneter aus Zelle a und ein Abgeordneter aus Zelle b.

(e) Die Untersuchungseinheiten können sowohl auf der X- als auch auf der Y-Dimension jeweils den gleichen Rangplatz einnehmen:

(4-16) $(X_A = X_B)$ und $(Y_A = Y_B)$

Diese (zweiseitig, doppelt) »in X und in Y verknüpften Paare« erhalten das Symbol $T_{xy}$. Wenn beide Abgeordnete der gleichen Zelle zugehören, bilden sie ein doppelt verknüpftes Paar.

Andere Paartypen sind nicht möglich.

Nach den Regeln der Kombinatorik (siehe Abschnitt 6.4, Teil II) lassen sich aus N Fällen N(N-1)/2 unterschiedliche Paare bilden, die man jeweils einem der fünf Paartypen zuordnen kann. In unserem Tabellenbeispiel erhalten wir folgende Paare (*Abb. 4.15*).

**Abb. 4.15:** Berechnung der Paartypen zu Abb. 4.14

| Paartyp | Rechengang | Summe |
|---|---|---|
| $N_c$ | a(e+f+h+i) + b(f+i) + d(h+i) + e(i) | 15175 |
| $N_d$ | c(d+e+g+h) + b(d+g) + f(g+h) + e(g) | 12678 |
| $T_x$ | a(d+g) + b(e+h) + c(f+i) + d(g) + e(h) + f(i) | 28081 |
| $T_y$ | a(b+c) + d(e+f) + g(h+i) + b(c) + e(f) + h(i) | 15503 |
| Txy | 1/2[a(a-1) + b(b-1) + ... + i(i-1)] | 15299 |
| | N(N-1)/2 = 417(416)/2 = | 86736 |

Es gibt unterschiedliche Möglichkeiten, diese Paar-Informationen zur Konstruktion von PRE-Maßzahlen zu nutzen. Drei von ihnen wollen wir in den folgenden Abschnitten vorstellen:

### 4.2.3.1 Die Maßzahl »Gamma«

Die Konstruktion dieser Kennzahl beruht ausschließlich auf Angaben über konkordante und diskordante Paare. Informationen über andere Paartypen werden ignoriert.

Überlegen wir zunächst, wie man die konkordanten und diskordanten Paare benutzen kann, um Prognoseregeln zu formulieren. Für unser konkretes Beispiel nehmen wir an, es bestehe ein positiver Zusammenhang

zwischen Schulbildung und Migrationsgrad: Abgeordnete mit einer höheren Schulbildung weisen tendenziell auch einen höheren Grad an Migration auf. Nehmen wir weiter an, für ein bestimmtes Abgeordnetenpaar (A,B) seien die Rangplätze der X-Variable bekannt und sie bildeten folgende Relation:

(4-17) $X_A < X_B$

Die Hypothese, es bestehe eine positive Beziehung zwischen den beiden Variablen, wird uns daraufhin zu der Prognose veranlassen,

(4-18) $Y_A < Y_B$

Wenn B auf der X-Variablen den höheren Rangplatz einnimmt, sagen wir voraus, daß B auch auf der Y-Variablen den höheren Rangwert aufweist. So werden wir bei beliebigen Paaren (A',B') verfahren, wenn sie auf der X-Dimension unterschiedliche Ränge einnehmen: Bei vermuteter positiver Beziehung prognostizieren wir für die Y-Variable die gleiche Relation wie die, die wir für $X_A$ und $X_B$ beobachtet haben. Wenn wir uns an diese Regel halten, machen wir immer dann einen Fehler, wenn nicht die gleichsinnige, sondern die zu X gegensinnige Relation in Y beobachtet wird, wenn das Paar also »diskordant« ist, da wir ja gemäß unserer Hypothese eines positiven Zusammenhangs stets die konkordante Relation prognostizieren. Insgesamt wird die Zahl unserer Fehler also $N_d$ sein. Fehler werden bei diesem Verfahren (bei dieser Kennzahl) nur als »vorgekommen« oder »nicht vorgekommen« gezählt. Es gibt keine »größeren« oder »kleineren« Fehler, da nicht registriert wird, um wieviel Rangplätze man sich verschätzt hat.

Wenn wir keinen positiven, sondern einen negativen Zusammenhang zwischen X und Y vermuten, ändern wir unsere Prognoseregel: Falls (4-17) gegeben ist, prognostizieren wir für Y nicht die Relation (4-18), sondern die gegensinnige Relation

(4-19) $Y_A > Y_B$

Jetzt machen wir eine falsche Voraussage gerade dann, wenn das Paar konkordant ist. Das bedeutet, unsere Fehlerzahl ist nun gleich $N_c$.

Aus all dem können wir folgenden Schluß ziehen: Wenn wir bei einer gegebenen bivariaten Verteilung die Kenntnis der unterschiedlichen X-Werte zweier Untersuchungseinheiten optimal im Sinne dieses Modells für die Prognose ihrer Werterelation auf der Y-Dimension verwenden, ist unsere Fehlermenge gleich dem Minimum von $N_c$ und $N_d$:

(4-20) $E_2 = \min(N_c, N_d)$

Jetzt benötigen wir noch die Vergleichsgröße $E_1$, die Menge der Fehler, die uns unterlaufen, wenn wir ohne Kenntnis der X-Rangplätze prognostizieren, welche der beiden Untersuchungseinheiten aus einem beliebigen Paar (A,B) den höheren (oder niedrigeren) Rangplatz auf der Y-Dimension einnimmt. Da wir nur von konkordanten und diskordanten Paaren ausgehen, muß eine der beiden Untersuchungseinheiten einen höheren Y-Rangplatz einnehmen als die andere. Wenn die Paare »gut gemischt« sind (bzw. zufällig ausgewählt werden),ist es gleichgültig, ob wir prognostizieren: das erstgenannte Paarelement hat den höheren Rangplatz, oder ob wir prognostizieren: das zweitgenannte Paarelement hat den höheren Rangplatz in Y. Auf jeden Fall werden wir 50 % richtige und 50 % falsche Prognosen abgeben. Folglich ist

(4-21) $E_1 = 0{,}5(N_c + N_d)$

Somit können wir den PRE-Koeffizienten gemäß (4-12) bilden:

$$(4\text{-}22) \qquad Y = \frac{E_1 - E_2}{E_1} = \frac{0{,}5(N_c+N_d) - \min(N_c,N_d)}{0{,}5(N_c+N_d)}$$

Wenn Zähler und Nenner mit 2 multipliziert werden, wird daraus

$$(4\text{-}22') \qquad Y = \frac{(N_c+N_d) - 2\min(N_c,N_d)}{(N_c+N_d)}$$

Wenn $N_c > N_d$, wenn also eine positive Beziehung vorliegt, erhalten wir

$$(4\text{-}22'') \qquad Y = \frac{(N_c + N_d - 2N_d)}{N_c + N_d} = \frac{N_c - N_d}{N_c + N_d}$$

Wenn $N_c < N_d$, wenn also eine negative Beziehung vorliegt, ergibt sich

$$(4\text{-}22''') \qquad Y = \frac{(N_c + N_d - 2N_c)}{N_c + N_d} = \frac{N_d - N_c}{N_c + N_d} = -\frac{N_c - N_d}{N_c + N_d}$$

In unserem Beispiel erhalten wir ein Gamma von

$$(4\text{-}23) \qquad \gamma = \frac{(N_c - N_d)}{N_c + N_d} = \frac{15175 - 12678}{15175 + 12678} = 0{,}09$$

Der Zusammenhang ist also nur ganz schwach ausgeprägt. Die Vorhersagefehler reduzieren sich nur um 9 %, wenn die Prognose auf der Kenntnis der X-Werte aufbaut. Wenn wir vorher einen stärkeren Zusammenhang erwartet haben, müssen wir nun nach möglichen Erklärungen für das überraschende Ergebnis suchen. Eine Erklärung könnte z. B. darin liegen, daß andere, bisher nicht berücksichtigte Faktoren den Einfluß der Schulbildung überlagern und modifizieren. Dieser Gedanke führt uns bereits in Richtung einer multivariaten Analyse, mit der wir uns erst im nächsten Kapitel eingehender beschäftigen wollen. Wir können aber einen bestimmten Aspekt schon an dieser Stelle einführen, ohne den Rahmen der bivariaten Tabellenanalyse auch im technischen Sinne zu überschreiten.

Sehen wir uns zunächst noch einmal die bivariate Tabelle in *Abb. 4.14 a* genauer an, nachdem wir die entsprechenden Prozentangaben hinzugefügt haben.

Der Gamma-Wert, der nur den **monotonen** Zusammenhang widerspiegelt, ist niedriger als der Betrag der formunspezifischen Maßzahl C = 0,10. Dies deutet darauf hin, daß der Zusammenhang eher kurvenförmig als monoton verläuft. Das wird erkennbar im Vergleich der Spaltenprozente (letzte Prozentangaben in jeder Zelle). Die Abgeordneten mit mittlerer Schulbildung weisen eher (zu 50 %) eine niedrigere Wanderungsintensität (1. Zeile) auf als die Abgeordneten mit niedriger **oder** hoher Schulbildung (39.8 bzw. 37.2 %). Hierzu korrespondieren die Spaltenprozente in der zweiten Zeile. Erst in der dritten Zeile (hohe Wanderungsintensität) steigen die Prozentanteile wie erwartet monoton von der ersten bis zur dritten Spalte leicht an.

Nicht-monotone (oder nicht-lineare) Beziehungen zwischen zwei Variablen sind häufig ein Hinweis darauf, daß ein weiterer Faktor wirksam ist, der identifiziert werden muß. In unserem Analysebeispiel liegt folgende Überlegung nahe: Unter den SPD-Abgeordneten befinden sich wahrscheinlich erheblich mehr Abgeordnete mit formal niedriger Schulbildung als unter den Parlamentariern bürgerlicher Parteien. (Das wäre also ein weiterer bivariater Zusammenhang.) Aber die SPD-Abgeordneten sind von ihrer Parteizentrale häufig in Wahlkreise außerhalb ihrer Heimatbezirke geschickt worden; politische und ökonomische Pressionen mögen zusätzliche Wanderschaften hervorgerufen haben. Somit entstand hier unabhängig von Bildungs- und entsprechenden Berufserfahrungen eine zu-

Abb.4.14a: Zusammenhang von Schulbildung und Wanderungsintensität (Reichstagsabgeordnete 1912)

| Wanderungsintensität | | Schulbildung: INIEDRIG (1) | MITTEL (3) | HOCH (5) | ROW TOTAL |
|---|---|---|---|---|---|
| NIEDRIG | 1 | 35 | 26 | 103 | 164 |
| | | 21.3 | 15.9 | 62.8 | 39.3 |
| | | 39.8 | 50.0 | 37.2 | |
| MITTEL | 2 | 38 | 15 | 113 | 166 |
| | | 22.9 | 9.0 | 68.1 | 39.8 |
| | | 43.2 | 28.8 | 40.8 | |
| HOCH | 3 | 15 | 11 | 61 | 87 |
| | | 17.2 | 12.6 | 70.1 | 20.9 |
| | | 17.0 | 21.2 | 22.0 | |
| COLUMN TOTAL | | 88 | 52 | 277 | 417 |
| | | 21.1 | 12.5 | 66.4 | 100.0 |

C = .10 V = .07 $\gamma$ = .09

sätzliche und relativ hohe Mobilität, die über derjenigen der bürgerlichen Abgeordneten liegen sollte. Wir können diese Überlegung indirekt überprüfen, indem wir den Zusammenhang zwischen Schulbildung und Wanderungsintensität für SPD-Abgeordnete und für Abgeordnete anderer Parteien getrennt untersuchen. (Wie das technisch mit SPSS$^x$ vor sich geht, zeigt das nächste Kapitel.) Hierzu betrachten wir die beiden Tabellen in *Abb. 4.14 b* und *4.14 c*.

Zunächst entnehmen wir den Randverteilungen, daß die »bürgerlichen« Abgeordneten in der Tat einen erheblich größeren Anteil (82.1 %) mit hoher formaler Bildung (Realgymnasium und höher) aufweisen als die SPD-Abgeordneten (25.9 %). Außerdem bestätigt sich, daß auf jeder Bildungsstufe die SPD-Abgeordneten einen höheren Anteil an mittlerer oder hoher Wanderungsintensität aufweisen als ihre bürgerlichen Kollegen. Auch der Zusammenhang zwischen Schulbildung und Wanderungsintensität tritt nun wesentlich stärker, aber auch differenzierter in Erscheinung. Bei den bürgerlichen Abgeordneten ist er mit einem Gamma = 0.45 indiziert, bei

Abb. 4.14b: Zusammenhang von Schulbildung und Wanderungsintensität bei SPD-Abgeordneten (1912)

| Wanderungsintensität | | Schulbildung: INIEDRIG (1) | MITTEL (3) | HOCH (5) | ROW TOTAL |
|---|---|---|---|---|---|
| NIEDRIG | 1 | 18 | 6 | 5 | 29 |
| | | 62.1 | 20.7 | 17.2 | 25.0 |
| | | 28.1 | 27.3 | 16.7 | |
| MITTEL | 2 | 32 | 10 | 15 | 57 |
| | | 56.1 | 17.5 | 26.3 | 49.1 |
| | | 50.0 | 45.5 | 50.0 | |
| HOCH | 3 | 14 | 6 | 10 | 30 |
| | | 46.7 | 20.0 | 33.3 | 25.9 |
| | | 21.9 | 27.3 | 33.3 | |
| COLUMN TOTAL | | 64 | 22 | 30 | 116 |
| | | 55.2 | 19.0 | 25.9 | 100.0 |

RAW CHI SQUARE = 2.26158
CRAMER'S V = .09873
CONTINGENCY COEFFICIENT = .13829
GAMMA = .18765

Abb. 4.14c: Zusammenhang von Schulbildung und Wanderungsintensität bei Abgeordneten außerhalb der SPD (1912)

| Wanderungsintensität | | Schulbildung: INIEDRIG (1) | MITTEL (3) | HOCH (5) | ROW TOTAL |
|---|---|---|---|---|---|
| NIEDRIG | 1 | 17 | 20 | 98 | 135 |
| | | 12.6 | 14.8 | 72.6 | 44.9 |
| | | 70.8 | 66.7 | 39.7 | |
| MITTEL | 2 | 6 | 5 | 98 | 109 |
| | | 5.5 | 4.6 | 89.9 | 36.2 |
| | | 25.0 | 16.7 | 39.7 | |
| HOCH | 3 | 1 | 5 | 51 | 57 |
| | | 1.8 | 8.8 | 89.5 | 18.9 |
| | | 4.2 | 16.7 | 20.6 | |
| COLUMN TOTAL | | 24 | 30 | 247 | 301 |
| | | 8.0 | 10.0 | 82.1 | 100.0 |

RAW CHI SQUARE = 16.31541
CRAMER'S V = .16463
CONTINGENCY COEFFICIENT = .22675
GAMMA = .44679

den SPD-Abgeordneten erwartungsgemäß schwächer mit einem Gamma = 0.19. In beiden Teiltabellen liegt der Koeffizient Gamma auch deutlich über den formunspezifischen Maßzahlen Cramers V und Kontingenzkoeffizient C. In Kapitel 5 werden wir solche Kausalkonstellationen unter den Stichworten »Interaktion« und »Suppression« näher erläutern. Zunächst aber wollen wir die formale Erörterung von Gamma und anderen Maßzahlen noch etwas weiter treiben.

Angewandt auf die 4-Felder-Tafel ist Gamma identisch mit **Yules Q:**

$$(4\text{-}24) \qquad Q = \frac{(N_c - N_d)}{N_c + N_d} = \frac{ad - bc}{ad + bc}$$

Die Buchstaben bezeichnen hier die vier beobachteten Zellenhäufigkeiten (a kennzeichnet die linke obere, d die rechte untere Zelle).

In (2 x 2)-Tabellen ist Gamma = Q nicht abhängig von den Randhäufigkeiten. Bei größeren Tabellen reagiert Gamma allerdings auf Veränderungen in den Randhäufigkeiten. Gamma reagiert auch, wenn die Zahl der Variablenausprägungen durch Zusammenlegen von Kategorien vermindert wird: es wird (unter sonst gleichen Bedingungen) größer und zwar in höherem Ausmaß als die in den beiden folgenden Abschnitten besprochenen Koeffizienten von Kendall und Somers (siehe Beispiele in Benninghaus 1976: 163 ff.). Diese Eigenschaft ist vor allem deshalb beklagenswert, weil die Zahl der Kategorien vom Forscher oft willkürlich festgelegt wird, vor allem dann, wenn die theoretisch gemeinte Merkmalsdimension kontinuierlich ist, das Intervallskalenniveau aber lediglich aus meßtechnischen Gründen nicht realisiert werden kann. Selbst bei »natürlichen« Dichotomien (wie »Geschlecht«), die kreuztabelliert werden, führt Gamma zu Problemen, weil es den Höchstwert auch dann erreicht, wenn nur eine Zelle unbesetzt ist. Man wird aber kaum eine gemeinsame Verteilung wie die in *Abb. 4.16* als Indikator eines perfekten Zusammenhangs interpretieren wollen (für ein Gegenargument siehe Davis 1971:42).

Auch bei größeren Tabellen führen solche »Eckkorrelationen«, bei denen nur eine äußere Spalte und eine äußere Zeile besetzt sind, zum Höchstwert bei Gamma (im Unterschied zu Kendalls Tau und Somers' d) (siehe wiederum die Beispiele in Benninghaus 1976: 162). Dieses Problem entsteht dadurch, daß Gamma die Informationen über Verknüpfungen (»tied pairs«) ignoriert. »By treating much of the data (ties) as irrelevant, gamma achieves large and perhaps misleading measures of error reduction. For this reason, even though it appeared earlier in the literature and, as a consequence, has been used more widely than any of the d measures,

Abb. 4.16: Bivariate Verteilung
mit Gamma = 1

| | $x_1$ | $x_2$ | |
|---|---|---|---|
| $y_1$ | 50 | 0 | 50 |
| $y_2$ | 50 | 100 | 150 |
| | 100 | 100 | 200 |

gamma seems inadequate for evaluating 'The more X, the more Y'« (Hildebrand/Laing/Rosenthal 1977: 46).

#### 4.2.3.2 Kendalls Tau

Diese Maßzahl beruht ebenfalls auf einem Vergleich konkordanter und diskordanter Paare, berücksichtigt aber im Nenner des Quotienten auch die einseitig verknüpften Paare $T_x$ und $T_y$:

$$(4\text{-}25) \qquad \tau_b = \frac{(N_c - N_d)}{\sqrt{\left((N_c + N_d + T_x)\ (N_c + N_d + T_y)\right)}}$$

In unserem Beispiel (*Abb. 4.14*) ist $Tau_b = 0{,}05$, also kleiner als Gamma.

Für die gebundenen Paare (»Ties«) gibt es bei Tau eine Art Punktabzug. Das erscheint als sinnvoll, wenn man eine perfekte Beziehung nur dann als gegeben betrachtet, wenn jede Erhöhung des X-Wertes mit einer Erhöhung des Y-Wertes einhergeht (positive Beziehung) oder jede Erhöhung des X-Wertes mit einer Minderung des Y-Wertes verbunden ist (negative Beziehung). Tau(b) erreicht also den Höchstbetrag | 1 | nur in einer quadratischen Tabelle, wenn entweder ausschließlich die Zellen der Hauptdiagonalen oder ausschließlich die Zellen der Nebendiagonalen besetzt sind.

Man mag es aber als Nachteil empfinden, daß Tau(b) den Höchstwert nur in einer quadratischen Tabelle erreichen kann; schließlich will der Sozialforscher nicht ausschließlich Variablen mit einer gleichen Zahl von Kategorien kreuztabellieren. Deshalb hat Kendall (bzw. Stuart) eine Modifikation des Tau-Koeffizienten vorgeschlagen:

$$(4\text{-}26) \qquad \tau_c = \frac{(N_c - N_d)}{\frac{1}{2} N^2 \left(\frac{m-1}{m}\right)} \quad , \quad \text{wobei } m = \min(r,c)$$

Der Tau(c)-Koeffizient nimmt in einer rechteckigen Tabelle gegenüber der Tau(b)-Version im allgemeinen einen etwas höheren Betrag an. Hildebrand et al. 1977, S. 52 merken hierzu kritisch an: »the procedure appears ad hoc from the viewpoint of prediction analysis«. Im Gegensatz zu Tau(b) kann Tau(c) nicht im Sinne eines PRE-Maßes interpretiert werden (ebd.). Zur PRE-Interpretation von Tau(b) sei der Leser auf den Artikel von Wilson (1969) verwiesen.

Mit dem Namen Kendalls ist noch ein weiterer Tau-Koeffizient, $Tau_a$, verbunden:

$$(4\text{-}27) \qquad \tau_a = \frac{(N_c - N_d)}{\frac{N(N-1)}{2}}$$

Wir wollen ihn hier nicht weiter erörtern, da er nur dann aussagekräftig ist, wenn überhaupt keine Ties vorkommen, wenn also kein einziger Rangplatz der beteiligten Variablen von mehr als einer Untersuchungseinheit besetzt ist.

Für diesen Fall ist auch der sog. **Rangkorrelationskoeffizient** von Spearman definiert worden. Er läßt Anpassungskorrekturen zu, wenn die Zahl der »Ties« sehr gering ist. Er ist aber mit Vorsicht zu verwenden, da er die Rangplätze wie Werte von Intervallskalen behandelt.

Um die Verwirrung zu vergrößern, haben auch Goodman und Kruskal einen »Tau«-Koeffizienten konstruiert, der im Unterschied zu Kendalls Tau-Koeffizienten die Zusammenhangsstärke zwischen nominalen Variablen ausdrückt, übrigens auch mit einer PRE-Interpretation.

Anzumerken ist, daß in einer (2 x 2)-Tabelle mit 0/1-kodierten Variablen Tau(b) den gleichen Wert annimmt wie Phi und der Produkt-Moment-Korrelationskoeffizient von Pearson (siehe Aschn. 4.2.4).

#### 4.2.3.3 Somers' d-Koeffizient

Im Unterschied zu den symmetrischen Koeffizienten Gamma und Kendalls Tau hat Somers zwei asymmetrische Koeffizienten vorgeschlagen:

(4-28)
$$d_{yx} = \frac{(N_c - N_d)}{N_c + N_d + T_y}$$

$$d_{xy} = \frac{(N_c - N_d)}{N_c + N_d + T_x}$$

In unserem Beispiel (*Abb. 4.14*) ist $d_{yx} = 0{,}06$ und $d_{xy} = 0{,}045$.

Wenn Y als abhängige und X als unabhängige Variable angesehen werden können, scheint es sinnvoll, den Koeffizienten durch Einsetzen der nur in Y gebundenen Paare zu mindern (im Vergleich zu Gamma). Denn bei einer perfekten Beziehung, die von X nach Y verläuft, dürften keine Paare auftreten, die sich hinsichtlich ihrer X-Werte unterscheiden, nicht aber hinsichtlich ihrer Y-Werte. Entsprechendes gilt bei umgekehrter Kausalrichtigung. Hier sollten bei perfekter Beziehung keine Fälle auftreten, die sich in den Y-Werten unterscheiden, nicht aber in ihren X-Werten. Deshalb erscheinen diese einseitig gebundenen Paare im Nenner des Quotienten.

Somers (1968) hat für seine Maßzahl ebenfalls eine PRE-Interpretation ausgearbeitet.

In einer (2x2)-Tabelle ist Somers' $d_{yx}$ identisch mit der entsprechenden Prozentsatzdifferenz d%. Somers' d ist in der Literatur auch als Regressionskoeffizient (siehe Teil II, Kap. 10) für ordinale Variable bezeichnet worden (siehe z. B. Jarausch/Arminger/Thaller 1985, S. 166).

Allgemein gilt auch die Beziehung

(4-29)
$$d_{yx} \cdot d_{xy} = \tau_b^{\,2}$$

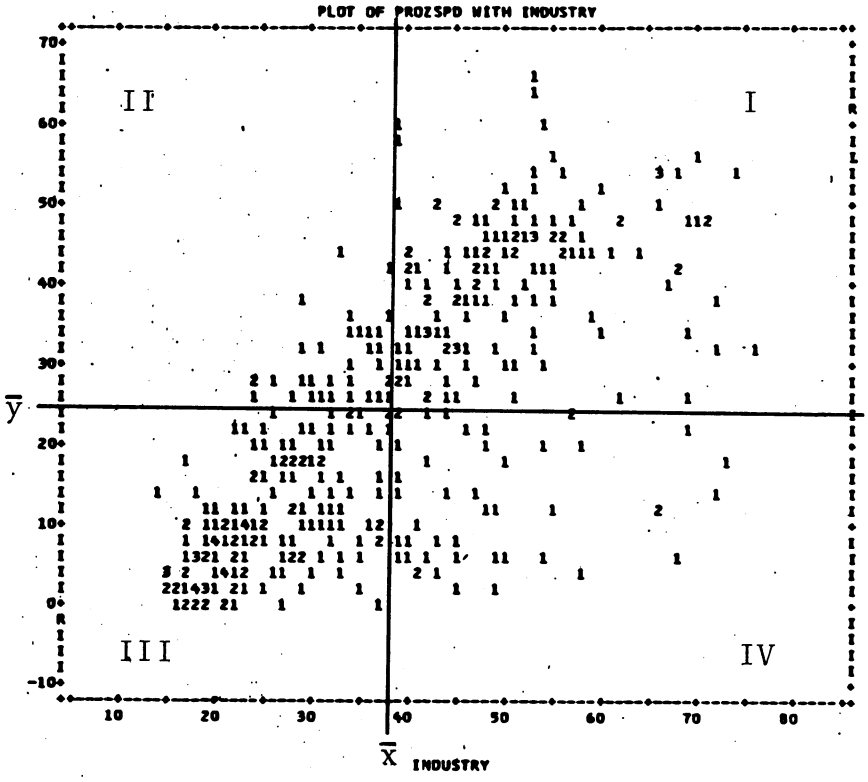
### 4.2.4 Ein Zusammenhangsmaß für metrische Variablen: Pearsons Produkt-Moment-Korrelationskoeffizient r

Die Ableitung dieses Korrelationskoeffizienten läßt sich am besten anhand eines Streudiagramms erläutern. Deshalb wiederholen wir an dieser Stelle die *Abb. 4.3* (als *Abb. 4.17*).

In das Koordinatenkreuz sind die Koordinaten der arithmetischen Mittel $\bar{y}$ und $\bar{x}$ neu eingezeichnet. Dadurch werden alle Punkte einer von vier »Regionen«, den sog. Quadranten, zugeteilt:

Quadrant I: alle ($y_i > \bar{y}, x_i > \bar{x}$)
Quadrant II: alle ($y_i > \bar{y}, x_i < \bar{x}$)
Quadrant III: alle ($y_i < \bar{y}, x_i < \bar{x}$)
Quadrant IV: alle ($y_i < \bar{y}, x_i > \bar{x}$)

Abb. 4.17: Prozentanteile der SPD (Ordinate) und Industrialisierungsgrad der Wahlkreise (Abszisse)




Mit dieser Flächeneinteilung soll folgender Gedankengang veranschaulicht werden: Wenn ein Zusammenhang zwischen zwei Variablen besteht, kann er sich in einer von zwei »Tendenzen« ausdrücken:

(1) Untersuchungseinheiten mit relativ großen (kleinen) X-Werten, haben auch relativ große (kleine) Y-Werte (positive Beziehung, entspricht einem Überwiegen konkordanter Paare bei Ordinalvariablen)

(2) Untersuchungseinheiten mit relativ großen (kleinen) X-Werten haben relativ kleine (große) Y-Werte (negative Beziehung, entspricht einem Überwiegen diskordanter Paare bei Ordinaldaten)

Als Kriterium für »relativ groß« oder »relativ klein« wollen wir den Abstand vom arithmetischen Mittel betrachten. Wir müssen also eine Maßzahl konstruieren, die

(a) bei positiver Beziehung einen um so größeren Wert annimmt, je größer der Anteil der Fälle im 1. und 3. Quadranten im Vergleich zum Anteil der Fälle im 2. und 4. Quadranten ist;
(b) bei negativer Beziehung einen um so niedrigeren Wert (größeren Absolutbetrag) annimmt, je größer der Anteil der Fälle im 2. oder 4. Quadranten im Vergleich zum Anteil der Fälle im 1. und 3. Quadranten ist.

Damit ist das Konzept der **Kovariation** formuliert. Sein formaler Ausdruck ist:

$$\text{(4-30)} \qquad \text{Kovariation } (y,x) = \sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Fälle im 1. und 3. Quadranten tragen positiv, Fälle im 2. und 4. Quadranten negativ zur Summe bei. Je gleichgewichtiger die Fälle über alle Quadranten gestreut sind, um so stärker nähert sich die Summe, die Kovariation, dem Wert Null.

Die Kovariation ist aber noch kein geeignetes Maß für die Zusammenhangsstärke. Offensichtlich ist sie von der Zahl der Fälle abhängig (es sei denn, es bestehe völlige Unabhängigkeit der Variablen, wodurch sich positive und negative Summanden in der Summe aufheben würden). Das kann man ändern, indem man die Kovariation durch die Zahl der Fälle dividiert:

$$\text{(4-31)} \qquad \frac{1}{N} \sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) = C_{yx}$$

Diesen Ausdruck bezeichnet man als **Kovarianz**. Er ist einer der Schlüsselbegriffe der Statistik. Aber als Maß für die bivariate Zusammenhangsstärke hat er noch Nachteile:

Offensichtlich hängt die Größe des Produkts in (4-31) jeweils davon ab, wieweit die beobachteten Werte $x_i$ und $y_i$ von $\bar{x}$ und $\bar{y}$ entfernt sind - gemessen in irgendwelchen Maßeinheiten. Das bedeutet, daß die Größe C von der jeweils verwendeten Skala abhängt. In unserem Beispiel erhalten wir also ganz unterschiedliche Ergebnisse, je nachdem, ob wir unsere Variablen in Prozent- oder in Promilleanteilen messen. Im ersten Falle ist $C_{yx} = 166{,}735$; im zweiten Falle ist $C_{yx} = 16673{,}5$ (siehe hierzu den Exkurs in Abschn 4.2.6 über das Rechnen mit Kovarianzen).

Ein Maß für die bivariate Zusammenhangsstärke sollte von den skaleninduzierten univariaten Streuungsumfängen möglichst unabhängig sein. Man kann diese Forderung wie folgt interpretieren: Der noch zu findende Korrelationskoeffizient sollte in seinem Absolutbetrag in dem Maße steigen, wie den **relativen** Abweichungen der einzelnen $x_i$-Werte vom arithmetischen Mittel $\bar{x}$ gleich große **relative** Abweichungen des jeweiligen $y_i$-Wertes von $\bar{y}$ korrespondieren. Unter der relativen Abweichung kann man das Verhältnis der Abweichung $(x_i - \bar{x})$ zur »durchschnittlichen« Abweichung, also zur Standardabweichung, verstehen:

$$(4\text{-}32) \qquad z(x)_i = \frac{x_i - \bar{x}}{s_x}$$

$$z(y)_i = \frac{y_i - \bar{y}}{s_y}$$

Man bezeichnet die gemäß (4-32) transformierten X- und Y-Variablen als **standardisierte** Variablen. Eine solche lineare Transformation ist, wie wir in Kap. 3 gesehen haben, für Intervallskalen zulässig. Das heißt, die internen Differenzenverhältnisse einzelner Wertepaare werden durch diese Transformation nicht berührt. Die standardisierten Variablen haben alle das arithmetische Mittel Null und eine Standardabweichung von Eins; denn es gilt allgemein (siehe die entsprechenden Ausführungen in Abschn. 3.1 und 3.2):

$$(4\text{-}33) \qquad \left\{ x^* = a + bx \right\} \quad \begin{array}{l} \rightarrow \overline{x^*} = b\bar{x} + a \\ \\ \rightarrow s_{x^*} = |b|\, s_x \end{array}$$

Bei der Standardisierung von Variablen gemäß (4-32) ist $b = 1/s$ und $a = (-\bar{x})/s$, folglich $\bar{z} = (1/s)\bar{x} - \bar{x}/s = 0$ und $s_z = (1/s)s = 1$.

Wenn wir die Kovarianzen nicht mit den beobachteten, sondern mit den z-transformierten Werten ausrechnen, folgt daraus

$$(4\text{-}34) \quad c_{z(y),z(x)} = \frac{1}{N} \sum \left(z(x)_i - \overline{z(x)}\right)\left(z(y)_i - \overline{z(y)}\right)$$

$$= \frac{1}{N} \sum z(x)_i \cdot z(y)_i$$

$$= \frac{1}{N} \sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \frac{1}{N} \cdot \frac{1}{s_x} \cdot \frac{1}{s_y} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{\frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N} \sum (x_i - \bar{x})^2} \sqrt{\frac{1}{N} \sum (y_i - \bar{y})^2}}$$

$$= \frac{\text{Kovarianz } (y,x)}{s_x s_y} = r$$

In unserem Beispiel (siehe *Abb. 4.17*) erhalten wir ein r = 0,63.

Auf die Rechentechnik gehen wir hier nicht ein, da uns der Computer die Arbeit abnimmt und die Rechenregeln in diesem Falle auch nicht helfen, das theoretische Korrelationskonzept hinter r zu verdeutlichen.

In SPSS$^x$ lautet das Standard-Kommando hierzu

```
PEARSON CORR SPDPROZ WITH INDUSTRY
STATISTICS 1, 2
```

Die beiden letzten Zeilen in (4-34) enthalten die »klassische« Formulierung des Pearsonschen Produkt-Moment-Korrelationskoeffizienten r. Sie zeigen, daß man die Variablen nicht vorgängig standardisieren muß, um r berechnen zu können. Man erhält das gleiche Ergebnis, wenn man die Kovarianz der ursprünglichen Variablen durch die beiden Standardabweichungen dividiert. Die Division durch die Standarabweichungen **beider** Variablen macht r zu einem symmetrischen Koeffizienten, d. h., für seine Berechnung spielt es keine Rolle, ob wir X oder Y als abhängige (oder unabhängige) Variable betrachten.

Der Koeffizient r mißt die Zusammenhangsstärke im Sinne einer **linearen** Beziehung zwischen den beiden Variablen. Das folgt daraus, daß lediglich die einfachen Differenzen zwischen beobachtetem Wert und arithmetischem Mittel in die Produktbildung eingehen und die Produktsumme anschließend nur noch linear transformiert wird. Das bedeutet auch, daß extreme Variablenwerte bei der Kovarianzberechnung ein stärkeres Gewicht erhalten als jene Werte, die nahe beim arithmetischen Mittel liegen.

Dies kann jedoch als wünschenswert angesehen werden. Wenn ein von der zentralen Tendenz der einen Variablen stark abweichender Wert verbunden ist mit einem ebenfalls extremen Wert der anderen Variablen, so mag dies als ein »gewichtigerer« Beleg für einen systematischen Zusammenhang der beiden Variablen gelten als eine Korrespondenz von X- und Y-Werten, die beide nahe bei »ihren« arithmetischen Mitteln liegen.

Der Korrelationskoeffizient r erreicht maximal den Betrag 1 (bei einer positiven Beziehung) und minimal den Betrag -1 (bei einer negativen Beziehung). Lineare Unabhängigkeit der beiden Variablen ist durch $r = 0$ angezeigt. (Ein nicht-linearer Zusammenhang kann dennoch bestehen.) Die Zwischenwerte sind nicht inhaltlich interpretierbar. Beim Vergleich zweier Koeffizienten kann man lediglich sagen, daß der größere r-Betrag einen stärkeren (linearen) Zusammenhang ausdrückt als der kleinere r-Betrag. Man kann aber, beispielsweise, nicht sagen, ein doppelt so großes $|r|$ indiziere einen doppelt so starken Zusammenhang. Der Koeffizient r kann jedoch im Rahmen des Regressionsmodells noch auf andere Weise abgeleitet werden. Die Größe $r^2$ wird dadurch inhaltlich interpretierbar und zwar wiederum als PRE-Maß (siehe hierzu Teil II, Kap. 10).

### 4.2.5 Zusammenhang zwischen einer nominalen und einer metrischen Variablen: Pearsons Eta

Bisher haben wir nur Beziehungen von Variablen untersucht, die beide das gleiche Meßniveau aufweisen. Sollte eine Variable mit höherem Meßniveau mit einer Variablen niedrigeren Meßniveaus kreuztabelliert werden, so müßte die Variable mit hohem Meßniveau nach den bisher besprochenen Analysemodellen auf das niedrigere Meßniveau der anderen Variablen herabgestuft werden. Wir wollen nun den Zusammenhang zwischen einer als »abhängig« gedachten metrischen Variablen und einer als »unabhängig« betrachteten Nominalvariablen untersuchen, dabei aber das metrische Meßniveau nicht herabstufen, sondern die metrischen Informationen nutzen.

Als Beispiel hierzu betrachten wir den Zusammenhang zwischen dem Lebensalter der Reichstagsabgeordneten und ihrer Fraktionszugehörigkeit. Neben der Restgruppe (»Sonstige«) werden fünf Fraktionen unterschieden: Sozialdemokraten, Linksliberale, Rechtsliberale, Zentrum, Konservative. Um die Ableitung des Zusammenhangsmaßes Eta leichter nachvollziehen zu können, gehen wir zunächst von einer nominalen Variablen aus, die nur zwei Kategorien, A und B, umfaßt. (Man kann sich hierunter natürlich wiederum zwei Fraktionen vorstellen.). Und wir begrenzen die Zahl der Untersuchungseinheiten auf 4 Fälle, die der Gruppe A angehören, und auf weitere 5 Fälle, die der Gruppe B angehören.

Die Daten lassen sich so anordnen, daß man zunächst das Lebensalter $y_i$ aller Untersuchungseinheiten i (i = 1,2,3,4) auflistet, die der Gruppe A angehören, danach die Lebensaltersangaben der Untersuchungseinheiten i (i = 5,6,...,9), die der Gruppe B angehören (siehe *Abb. 4.18*)

**Abb. 4.18:** Beispiel zur Berechnung von Eta

| Fall | $y_i$ | $y_i - \bar{y}_A$ | $(y_i - \bar{y}_A)^2$ | $y_i - \bar{y}_G$ | $(y_i - \bar{y}_G)^2$ | |
|---|---|---|---|---|---|---|
| 1 | 53 | 3,25 | 10,56 | -1,33 | 1,77 | |
| 2 | 61 | 11,25 | 126,56 | 6,66 | 44,35 | $\bar{y}_A$ = 49,75 |
| 3 | 38 | -11,75 | 138,06 | -16,33 | 266,67 | $s_A$ = 8,41 |
| 4 | 47 | -2,75 | 7,56 | -7,33 | 53,73 | |
| | | | 282,74 | | 383,49 | |
| | | $y_i - \bar{y}_B$ | $(y_i - \bar{y}_B)^2$ | | | |
| 5 | 69 | 11,0 | 121,0 | 14,66 | 214,92 | |
| 6 | 56 | -2,0 | 4,0 | 1,66 | 2,76 | |
| 7 | 49 | -9,0 | 81,0 | -5,33 | 28,41 | $\bar{y}_B$ = 58,0 |
| 8 | 54 | -4,0 | 16,0 | -0,33 | 0,11 | $s_B$ = 6,90 |
| 9 | 62 | 4,0 | 16,0 | 7,66 | 58,68 | |
| | | | 238,0 | | 304,88 | |
| | | | | | 688,37 | |
| | | $\bar{y}_G$ = 54,33, | $s_G$ = 8,75 | | | |

Unsere Frage nach dem Zusammenhang zwischen der Gruppenzugehörigkeit und dem Lebensalter können wir wiederum im Sinne der Prognosefähigkeit interpretieren: Wenn wir wissen, ob eine Person i Mitglied der Gruppe A (z. B. ein SPD-Mitglied) oder Mitglied der Gruppe B (z. B. der konservativen Fraktion) ist, sollten wir ihr Lebensalter im Falle des Zusammenhangs der beiden Variablen besser (d. h. mit einem geringeren Fehler) prognostizieren können als ohne diese Information. Also müssen wir zunächst wieder festlegen, wie der Prognosefehler zu berechnen ist. Bei metrischen Daten bietet sich hierfür die Differenz zwischen »wahrem« und prognostiziertem Wert an. Da wir die Fehler summieren wollen, quadrieren wir sie, so daß sich positive und negative Beträge nicht ausgleichen. Wir wissen aus Kap. 3, daß die Summe der Differenzen $(y_i - \bar{y})^2$ ein Minimum ergibt. Wenn wir die Y-Werte ohne Kenntnis der X-Werte (hier: ohne Kenntnis der Gruppenzugehörigkeit) prognostizieren, erhalten

wir die geringste Fehlerquadratsumme also dann, wenn wir in jedem einzelnen Fall stets den Mittelwert der Y-Verteilung als Prognosewert verwenden. Laut *Abb. 4.18* ist das

$$(4\text{-}35) \qquad \bar{Y}_G = \frac{1}{9}\,(\bar{Y}_A \cdot 4 + \bar{Y}_B \cdot 5) = 54.33$$

(Der Index G steht hier für die Gesamtheit der Fälle)

Daraus ergibt sich bei diesem ersten Prognoseverfahren eine Fehlerquadratsumme von

$$(4\text{-}36) \qquad E_1 = \sum_{i=1}^{9} (y_i - \bar{y}_G)^2 = 688{,}37$$

(siehe wiederum *Abb. 4.18*). Wenn wir bei einem zweiten Prognosedurchgang die Gruppenzugehörigkeit der einzelnen Personen kennen, prognostizieren wir nicht mehr $\hat{y}_i = \bar{y}_G$, sondern $\hat{y}_i = \bar{y}_A$, falls die betreffende Person der Gruppe A angehört, und $\hat{y}_i = \bar{y}_B$, falls sie zur Gruppe B gehört. Als Prognosefehler erhalten wir diesmal

$$(4\text{-}37) \qquad E_2 = \sum_{i=1}^{4} (y_i - \bar{y}_A)^2 + \sum_{i=5}^{9} (y_i - \bar{y}_B)^2 = 520{,}74$$

Somit können wir die proportionale Fehlerreduktion (PRE) als Indikator für die Zusammenhangsstärke wie folgt angeben:

$$(4\text{-}38) \quad PRE = \frac{E_1 - E_2}{E_1} = \frac{\sum_{i=1}^{9}(Y_i-\bar{Y}_B)^2 - \left[\sum_{i=1}^{4}(Y_i-\bar{Y}_A)^2 + \sum_{i=5}^{9}(Y_i-\bar{Y}_B)^2\right]}{\sum_{i=1}^{9}(Y_i-\bar{Y}_G)^2}$$

Um diesen Ausdruck noch besser interpretieren zu können, formen wir ihn um. Dazu führen wir neben dem Personenindex i einen zweiten Index, j, (j = 1,2,...,k) ein, der bei jeder Person die Gruppenzugehörigkeit angeben soll. In unserem bisherigen Zahlenbeispiel ist k = 2; j = 1 steht für Gruppe A und j = 2 steht für Gruppe B. Außerdem numerieren wir die Personen in den beiden Gruppen getrennt. Der Index i läuft nun also von 1 bis $n_j$ = Zahl der Fälle in der Gruppe j. Somit ist

$$(4\text{-}39) \qquad E_1 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_G)^2$$

$$E_2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$$

Das doppelte Summenzeichen bedeutet: Wir addieren die Differenzen über alle Personen i aus allen Gruppen j. Man hält zunächst den Gruppenindex j = 1 fest und addiert über alle $n_1$ Personen dieser ersten Gruppe. Dann erhöht man den Gruppenindex auf j = 2 und addiert über alle $n_2$ Personen. (das Rechnen mit Summenzeichen ist im Anhang erläutert).

Die Differenz des Beobachtungswertes $y_{ij}$ zum Prognosewert $\bar{y}_G$ in $E_1$ läßt sich in zwei Komponenten zerlegen: den Abstand von $y_{ij}$ zum Gruppenmittelwert $\bar{y}_j$ und den Abstand des Gruppenmittelwertes zum arithmetischen Mittel der Gesamtheit:

$$(4\text{-}40) \qquad Y_{ij} - \bar{Y}_G = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \bar{Y}_G)$$

Da $\bar{y}_j$ einmal mit positivem und einmal mit negativem Vorzeichen auftritt, wird die Summe nicht verändert. Wir benötigen die neue Schreibweise, um folgendes zeigen zu können:

Durch Quadrieren, Ausmultiplizieren des Binoms auf der rechten Gleichungsseite und Summieren über alle Fälle in allen Gruppen erhält man

$$(4\text{-}41) \quad \underbrace{E_1 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_G)^2}_{\text{Gesamtvariation}} = \underbrace{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}_{\text{nicht erklärte Variation}} + \underbrace{\sum_{j=1}^{k} (\bar{Y}_j - \bar{Y}_G)^2 \, n_j}_{\text{erklärte Variation}}$$

Die Binomkomponente $\Sigma\Sigma\,[\,2(\bar{y}_j - \bar{y}_G)\,(y_{ij} - \bar{y}_j)\,]$ wurde weggelassen, da sie Null ergibt: Für jede Gruppe ist die Komponente $(\bar{y}_j - \bar{y}_G)$ für alle Gruppenmitglieder eine Konstante; sie kann also vor das Summenzeichen gezogen werden. Die Summe $\Sigma\,(y_{ij} - \bar{y}_j)$ wird aber in jeder Gruppe Null, so daß das ganze Produkt Null wird.

Wir können somit nicht nur die einfachen Distanzen in zwei Komponenten zerlegen, sondern auch die »Variationen« oder »Varianzen«. (Die Varianzen erhält man bekanntlich, indem man die Variationen durch die Zahl der Fälle dividiert.) Daran können wir folgende Überlegung knüpfen: Wenn die Y-Variable vollständig durch die X-Variable »determiniert« wäre, dürfte die Variation der Y-Werte nur durch eine Variation der X-Werte ausgelöst werden. Die Variation der Y-Werte müßte, wie man sagt,

vollständig »zurückführbar«, »erklärbar« sein durch die Variation der X-Werte.

In den Sozialwissenschaften lassen sich in der Regel keine Experimente durchführen. Wir können die Veränderungen der Variablenwerte gleichsam nur »simulieren«, indem wir bei Fällen mit unterschiedlichen X-Werten prüfen, ob konsistent gleichsinnige oder konsistent gegensinnige Unterschiede auch in den Y-Werten feststellbar sind. (Selbst wenn wir Konsistenz beobachten, dürfen wir nicht sicher sein, daß wirklich eine kausale Beziehung vorliegt, da andere, nicht kontrollierte Einflußfaktoren den statistischen Zusammenhang zwischen den beiden beobachteten Variablen herbeigeführt oder beeinflußt haben können. (Hierzu mehr in Kap. 5.) Diese Überlegungen bleiben auch dann gültig, wenn wir nicht von einer asymmetrischen Kausalbeziehung, sondern nur von »irgendeiner« Art Zusammenhang sprechen.

In unserem Beispiel hat X nur wenige Ausprägungen, die Gruppenzugehörigkeiten ausdrücken. Bezogen auf diesen Fall bedeutet vollständiger Zusammenhang, daß **innerhalb** jeder Gruppe die Variation von Y gleich Null ist, daß aber die Y-Werte **zwischen** den Gruppen variieren. Alle Mitglieder der SPD-Fraktion müßten also das gleiche Lebensalter aufweisen, und alle Konservative müßten das untereinander auch. Mit anderen Worten: Ihre Y-Varianz müßte sich innerhalb jeder Gruppe auf Null »reduzieren«. Aber jeder SPD-Abgeordnete müßte ein anderes Alter haben als sein konservativer Kollege. Einen vollständigen Zusammenhang wird man nicht nur in diesem Beispiel, sondern auch in den Sozialwissenschaften generell so gut wie nie beobachten können. Aber eine Grenzfallbetrachtung wie diese läßt die Konzeption des Zusammenhangsbegriffs schärfer hervortreten. Wenn ein perfekter Zusammenhang durch vollständige »Varianzreduktion« im eben erläuterten Sinne charakterisiert werden kann, dann lassen sich unterschiedliche Grade der Zusammenhangsstärke durch den **Anteil** der reduzierten Variation an der Gesamtvariation numerisch schlüssig kennzeichnen.

Hierzu betrachten wir erneut die Komponenten der Variation, wie sie in Gleichung (4-41) auftreten. Der erste Ausdruck auf der rechten Seite bezeichnet die Variation in Y **innerhalb** jeder Gruppe. Das ist also derjenige Anteil an der Gesamtvariation, der **nicht** durch die Gruppenvariable »erklärt« werden kann (denn die Fälle einer Gruppe haben ja alle den gleichen X-Wert). Der zweite Ausdruck bezeichnet denjenigen Variationsanteil, den wir durch die Gruppenvariable erklären können. Er enthält nämlich die Summe der quadrierten Prognosefehler, die uns **weniger** unterlaufen, wenn wir statt des Gesamtmittels $\bar{y}_G$ das Gruppenmittel $\bar{y}_j$ als Prognosewert benutzen. Dadurch kommen wir (falls überhaupt ein Zusammenhang besteht) bei unserer Prognose im Durchschnitt näher an den wahren Wert heran; im Durchschnitt erreichen wir bei jedem Wert eine Näherung um die Strecke $\bar{y}_j - \bar{y}_G$.

Wir stellen nun die Gleichung (4 - 41) um, indem wir den ersten Ausdruck der rechten Seite von beiden Gleichungsseiten subtrahieren und die neue Gleichung durch den ersten Ausdruck dividieren:

$$(4\text{-}42)\quad \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_G)^2 - \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_j)^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_G)^2} = \frac{\sum_{j=1}^{k}(\bar{Y}_j-\bar{Y}_G)^2\, n_j}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_G)^2}$$

$$= \eta^2$$

Der Ausdruck auf der linken Seite ist offensichtlich identisch (siehe Gleichungen (4-38) und (4-39)) mit unserem PRE-Maß $(E_1 - E_2)/E_1$, das in dieser Gestalt in der Literatur mit dem Symbol $\eta^2$ = Eta$^2$ bezeichnet wird.

Dieser Korrelationskoeffizient, der auch als »correlation ratio« geführt wird, drückt den Anteil der erklärten Variation an der Gesamtvariation aus. Er kann folglich nur Werte zwischen 0 und 1 annehmen. (Aus naheliegenden Gründen ersparen wir es uns wiederum, die günstigste Methode zur Berechnung von Eta$^2$ zu erläutern.)

Wir präsentieren nun noch (in *Abb. 4.19*) den Computerausdruck, der das Ergebnis aus der realen Analyse des Zusammenhangs zwischen dem Alter der Reichstagsabgeordneten und ihrer Fraktionszugehörigkeit wiedergibt. Wir erhalten ihn in SPSS$^x$ mit folgendem Befehl:

```
BREAKDOWN ALTER BY FRAKTION
STATISTIC    1
```

Er zeigt, daß nur knapp 7 % der Altersvariation der Abgeordneten durch die Fraktionszugehörigkeit »erklärbar« ist. Der Zusammenhang ist also recht schwach.

Die vorangegangene Diskussion sollte deutlich gemacht haben, daß der Begrif der Zusammenhangsstärke nicht eindeutig ist. Er enthält zwei Komponenten, die nicht aufeinander zurückführbar sind:

(1) den Grad der Zuverlässigkeit oder Sicherheit, mit der eine Änderung der X-Werte mit einer Änderung der Y-Werte verbunden ist, wobei das quantiative Ausmaß dieser Änderung offenbleibt. Diese Komponente kann durch das Konzept der Varianzreduktion oder der »Bindung« der Varianz in Y durch die Varianz in X operationalisiert werden.

(2) das Ausmaß, die »Strecke«, um die sich Y im Durchschnitt verändert, wenn sich X um eine bestimmte Einheit ändert. Diese Strecke läßt sich in diesem Falle durch die Größe der Mittelwertdifferenz

Abb. 4.19: Ergebnisausdruck zur Berechnung von Eta

CRITERION VARIABLE ALT

ANALYSIS OF VARIANCE

| VARIABLE | CODE | VALUE LABEL | SUM | MEAN | STD DEV | SUM OF SQ | N |
|---|---|---|---|---|---|---|---|
| V77 | 1 | SOZDEM | 5113.0000 | 41.9098 | 6.6854 | 5408.0082 | ( 122) |
| V77 | 2 | LIBD | 2211.0000 | 45.1224 | 9.2750 | 4129.2653 | ( 49) |
| V77 | 3 | RELIB | 2523.0000 | 45.8727 | 9.0575 | 4430.1091 | ( 55) |
| V77 | 4 | ZENTRUM | 4828.0000 | 43.4955 | 7.7622 | 6627.7477 | ( 111) |
| V77 | 5 | KONSERVATIV | 3438.0000 | 47.7500 | 6.9358 | 3415.5000 | ( 72) |
| | | TOTAL | 18113.0000 | 44.2861 | 7.9529 | 25805.5306 | ( 409) |

ANOVA TABLE

| | SUM OF SQUARES | DEGREES OF FREEDOM | MEAN SQUARE |
|---|---|---|---|
| BETWEEN GROUPS | 1794.9002 | ( 4) | 448.7251 |
| WITHIN GROUPS | 24010.6303 | ( 404) | 59.4323 |
| TOTAL | 25805.5306 | ( 408) | |

F = 7.5502 SIG. = .0000 ETA SQRD = .0696

zweier Gruppen ausdrücken. Ein hoher Eta-Koeffizient wäre durchaus vereinbar mit einer sehr geringen Mittelwertdifferenz.

Der Sozialforscher, nicht der Statistiker muß entscheiden, welche der beiden Komponenten ihm bei einer Forschungsfrage die wichtigste ist. In der Regressionsanalyse (Kap. 10) werden beide Komponenten aus einem integrierten Ansatz heraus entwickelt und in zwei Koeffizienten numerisch spezifiziert.

Zum Schluß noch drei Hinweise:

(a) $Eta^2$ wird tendenziell um so größer, je mehr Kategorien die Nominalvariable umfaßt, je mehr Gruppen also definiert sind.

(b) Im Unterschied zu den anderen Kennzahlen bivariater Verteilungen, die wir bisher besprochen haben, kann $Eta^2$ nicht direkt statistisch getestet werden (siehe Kap. 8), sondern nur nach einer Modifikation (siehe Bortz 1979: 343 f.). In der Regel benutzt man aber in solchen Fällen den varianzanalytischen F-Test.

(c) $Eta^2$ wird in der Regel nur im Kontext eines umfassenderen Analysedesigns, der sog. Varianzanalyse, verwendet, die wir hier aber nicht weiter erörtern. Allerdings läßt sich die Varianzanalyse als Sonderfall der Regressionsanalyse auffassen, die wir in Teil II darstellen. Dabei spielt das eben erläuterte Konzept der Varianzzerlegung eine wichtige Rolle.

### 4.2.6 Exkurs: Das Rechnen mit Kovarianzen[3] (*)

In Lehrbüchern zur »fortgeschrittenen« Statistik und Datenanalyse, aber auch in Forschungsberichten und Zeitschriftenartikeln wird häufig mit Kovarianzen »gerechnet«, ohne daß die entsprechenden algebraischen Regeln vorgestellt werden. Ihre Kenntnis wird schlicht vorausgesetzt, obwohl die einführenden Texte zur statistischen Datenanalyse sie nur selten erläutern. Wir wollen die wichtigsten Regeln und Ableitungen ohne ausführliche Beweisführung hier wenigstens nennen, um eine spätere Lektüre fortgeschrittener Texte in diesem Punkte zu erleichtern.

Zunächst wiederholen wir die Definitionen für Varianzen (V) und Kovarianzen (C) aus vorangegangenen Abschnitten:

$$(4\text{-}43) \qquad V(x) = \frac{1}{N} \sum (x_i - \bar{x})^2$$

$$C(X,Y) = \frac{1}{N} \sum (x_i - \bar{x})\ (y_i - \bar{y})$$

[3] Vergl. zum folgenden Kenny 1979, S. 17 ff.

(Wir sparen uns wiederum die inferenzstatistische Korrektur des Nenners von N zu N-1).

Die Varianz läßt sich mathematisch als »Autokovarianz«, als Kovarianz einer Variablen mit sich selbst, verstehen. So werden wir frühere Aussagen über Varianz und Standardabweichungen (siehe Gleichungen (3-10) und (3-11)) in den folgenden Regeln wiederfinden.

**Regel 1**

(4 - 43a) $C(X,k) = 0$

folgt unmittelbar aus der Definition (4 - 43): Die Kovarianz einer Variablen X mit einer Konstanten Y = k ist Null. Da k für alle Objekte i, i = 1,2,...,N, den gleichen Betrag hat, ist das arithmetische Mittel $\bar{y} = \bar{k} = k$ und somit $y_i - \bar{y} = k - k = 0$. Folglich wird das Produkt $(x_i - \bar{x})(y_i - \bar{y}) = 0$.

**Regel 2**

(4 - 44) $C(aX,bY) = a \cdot b \cdot C(X,Y)$

Werden die Variablen X und Y mit dem Faktor a bzw. b multipliziert, verändert sich die Kovarianz um den Faktor a·b.
Falls a = 1 und b = - 1, gilt C(X, - Y) = - C(X,Y).
Diese Regel läßt sich auf die Varianz anwenden, indem man Y = X und a = b = k setzt:

(4 - 44a) $C(kX,kX) = k^2C(X,X) = k^2V(X)$.

**Regel 3**

(4 - 45) $C(X, Y+Z) = C(X,Y) + C(X,Z)$

Die Kovarianz einer Variablen X mit der Summe zweier Variablen, Y und Z, ist gleich der Summe der Kovarianz dieser Variablen X mit jeder der Komponenten der Summe.

Aus der Summenregel (4 - 45) folgt unmittelbar

(4 - 46) $C(aX+bY, X+cZ) = C(aX+bY, X) + C(aX+bY, cZ)$,

wobei zunächst (aX + bY) als erste und (X + cZ) als zweite (Summen-)Variable behandelt werden. In den beiden Klammerausdrücken auf der rechten Gleichungsseite steht nun jeweils eine Variable, X bzw. cZ, die erneut mit einer Summenvariable jeweils »gekoppelt« ist. Eine neuerliche Anwendung der Summenregel führt somit zu

(4 - 46a) $= C(aX,X) + C(bY,X) + C(aX,cZ) + C(bY,cZ)$

Darauf läßt sich (4-44), die Konstantenregel, anwenden:

(4 - 46b) $= aC(X,X) + bC(Y,X) + a \cdot c \cdot C(X,Z) + b \cdot c \cdot C(Y,Z),$

wobei der erste Ausdruck definitionsgemäß [siehe (4 - 43)] zu aV(X) zu verkürzen wäre. Man beachte, daß in den Klammern die Elemente der Kreuzprodukte aus $(X+Y)(X+Z) = XX + YX + XZ + YZ$ stehen.

Ein sehr wichtiges Theorem ist

**Regel 4**

(4 - 47) $V(aX + bY) = a^2[V(X)] + b^2[V(Y)] + 2ab[C(X,Y)]$

Wenn $a = b = 1$, wird daraus:

(4 - 47a) $V(X + Y) = V(X) + V(Y) + 2C(X,Y)$

Die Varianz der Summe zweier Variablen ist gleich der Summe der Varianzen dieser beiden Variablen plus des Zweifachen ihrer Kovarianz. Handelt es sich nicht um die Summe, sondern um die Differenz zweier Variablen, wird, wegen $b = -1$ die doppelte Kovarianz nicht addiert, sondern subtrahiert.

Dieses Theorem folgt aus der Definition der Varianz - $V(X+Y) = C(X+Y,X+Y)$ - und der Anwendung der Summenregel (Regel 3):

(4 - 48) $C(X+Y, X+Y) = C(X,X) + C(X,Y) + C(Y,X) + C(Y,Y)$

Da die Kovarianz symmetrisch ist, $C(X,Y) = C(Y,X)$, folgt (4 - 47a) aus (4 - 48) nach Anwendung der Varianzdefinition. Das Theorem (4 - 47a) läßt sich auf mehr als zwei Variablen verallgemeinern: Die Varianz einer Summe ist gleich der Summe der Varianzen der einzelnen Variablen plus dem Zweifachen aller möglichen Kovarianzen, die sich aus diesen Variablen bilden lassen.

### 4.2.7 Exkurs: Eingeschränkte Variation der abhängigen Variablen (*)

Kehren wir noch einmal zu *Abb. 4.14 b* zurück, die den Zusammenhang zwischen formaler Bildung und Wanderungsintensität der SPD-Reichstagsabgeordneten von 1912 darstellt (hier wiederholt als *Abb. 4.20*)

Nehmen wir an, aus irgendeinem Grunde wären die Unterlagen für die Abgeordneten mit niedriger Schulbildung verloren gegangen. Dann erhielten wir eine neue Tabelle, die sich von der bisherigen nur durch das Fehlen der ersten Spalte unterschiede. Nicht nur die absoluten, sondern auch die relativen Häufigkeiten in den verbleibenden Zellen blieben unverän-

Abb. 4.20: Zusammenhang von Schulbildung und Wanderungsintensität bei SPD-Abgeordneten (1912)

| Wanderungsintensität | | Schulbildung: INIEDRIG | MITTEL | HOCH | ROW TOTAL |
|---|---|---|---|---|---|
| | | 1. | 3. | 5. | |
| NIEDRIG | 1. | 18 | 6 | 5 | 29 |
| | COL PCT | 28.1 | 27.3 | 16.7 | |
| | ROW PCT | 62.1 | 20.7 | 17.2 | 25.0 |
| MITTEL | 2. | 32 | 10 | 15 | 57 |
| | COL PCT | 50.0 | 45.5 | 50.0 | |
| | ROW PCT | 56.1 | 17.5 | 26.3 | 49.1 |
| HOCH | 3. | 14 | 6 | 10 | 30 |
| | COL PCT | 21.9 | 27.3 | 33.3 | |
| | ROW PCT | 46.7 | 20.0 | 33.3 | 25.9 |
| COLUMN TOTAL | | 64 | 22 | 30 | 116 |
| | | 55.2 | 19.0 | 25.9 | 100.0 |

RAW CHI SQUARE = 2.26158 WITH
CRAMER'S V = .09873
CONTINGENCY COEFFICIENT = .13829
GAMMA = .18765

Abb. 4.21: Zusammenhang von Schulbildung und Wanderungsintensität bei SPD-Reichstagsabgeordneten (1912) nach fiktivem Verlust der Fälle mit ranghöchster Wanderungsintensität

| Wanderungsintensität | | Schulbildung: INIEDRIG | MITTEL | HOCH | ROW TOTAL |
|---|---|---|---|---|---|
| | | 1. | 3. | 5. | |
| NIEDRIG | 1. | 18 | 6 | 5 | 29 |
| | COL PCT | 36.0 | 37.5 | 25.0 | 33.7 |
| MITTEL | 2. | 32 | 10 | 15 | 57 |
| | COL PCT | 64.0 | 62.5 | 75.0 | 66.3 |
| COLUMN TOTAL | | 50 | 16 | 20 | 86 |
| | | 58.1 | 18.6 | 23.3 | 100.0 |

RAW CHI SQUARE = .89902 WITH
CRAMER'S V = .10224
CONTINGENCY COEFFICIENT = .10171
GAMMA = .14650

dert, insbesondere blieben die Prozentdifferenzen (Spaltenprozente) zwischen den Abgeordneten mit mittlerer und den Abgeordneten mit hoher Schulbildung stabil. Wir müßten lediglich unsere Aussage: »Bei Abgeordneten mit höherer Schulbildung zeigt sich eine etwas stärkere Wanderungsintensität als bei Abgeordneten mit niedrigerer Schulbildung« auf den tatsächlich untersuchten Personenkreis einschränken, also anmerken: Dies ist nur für Abgeordnete nachgewiesen, die mindestens ein mittleres Niveau formaler Schulbildung erreicht haben.

Nehmen wir nun einen zweiten Fall des Datenverlusts an, in dem aus irgendwelchen Gründen keine Angaben von SPD-Abgeordneten mit hoher Wanderungsintensität vorliegen. Dann erhalten wir die Tabelle in *Abb. 4.21*.

In der neuen Tabelle entfällt nun gegenüber der alten nicht eine Spalte, sondern eine Zeile. Wenn die Zeilenvariable die abhängige Variable ist, wird durch diesen »Schnitt« der Aussagegehalt der bivariaten Verteilung verändert, denn zwischen den bedingten Verteilungen ergeben sich nun (von Spalte zu Spalte) andere (Prozent-)Differenzen. So ist z. B. nicht mehr erkennbar, daß SPD-Abgeordnete mit mittlerer Schulbildung eine höhere Wanderungsintensität aufweisen als SPD-Abgeordnete mit niedrigerer Schulbildung. Die Prozentdifferenz zwischen der mittleren und der hohen Bildungsstufe behält zwar ihr Vorzeichen, verändert aber ihren Betrag. Das bedeutet: Wenn die vollständige Tabelle in *Abb. 4.20* zu gültigen Ergebnissen führt, gelangen wir mit der unvollständigen Tabelle in *Abb. 4.21* zu einer nicht-validen »Schätzung« der Prozentdifferenzen. Dieser Defekt ist auch nicht dadurch korrigierbar, daß man die Aussage auf Abgeordnete begrenzt, die lediglich eine niedrigere oder mittlere Wanderungsintensität aufweisen. Im vorliegenden Falle mag die Verzerrung substantiell unerheblich sein. Doch beleuchtet dieses simple Beispiel ein wichtiges Problem, das gerade für die historische Sozialforschung besonders relevant ist, da sie es häufig mit unvollständigen Quellen zu tun hat. Solange diese Quellendefekte lediglich den Wertebereich bzw. die Varianz der unabhängigen Variablen X einschränken, nicht aber die Variation der abhängigen Variable Y bei gegebenen X-Werten, bleiben gültige Aussagen über den »strukturellen« Zusammenhang der beiden Variablen möglich, sofern die Aussage auf den tatsächlich erhobenen Wertebereich in X eingeschränkt wird. (Der Begriff des strukturellen Zusammenhangs wird in Teil II, Kap. 10 näher verdeutlicht.) Wird dagegen die Variation der Y-Werte begrenzt, sind insbesondere bestimmte Wertebereiche oberhalb oder unterhalb eines »Schwellenwertes« regelrecht »abgeschnitten«, so sind selbst diejenigen Strukturkoeffizienten (wie Prozentsatzdifferenzen) nicht mehr valide, die für den begrenzten Y-Wertebereich ermittelt wurden.

Es ist z. B. nicht (oder nur mit Hilfe zusätzlicher Modellannahmen) möglich, zu gültigen Aussagen über den Zusammenhang zwischen sozialstrukturellen Faktoren (wie Industrialisierungsgrad) und SPD-Stimmenanteilen zu gelangen, wenn für die Untersuchung lediglich Wahlbezirke mit einem SPD-Stimmenanteil von mindestens 15 % (geplant oder ungeplant) ausgewählt wurden. Auch kann man, um ein anderes Beispiel zu erwähnen, den Einfluß bestimmter Faktoren auf die Wahrscheinlichkeit des Auftretens krimineller Handlungen bei Personen oder Kollektiven nicht zuverlässig ermitteln, wenn nur solche Untersuchungseinheiten erhoben wurden, bei denen lediglich wenige ausgewählte Typen krimineller Handlungen zu beobachten sind. Eine genauere Erörterung dieser Problematik (siehe z. B. Berk 1983; Berk/Ray 1982) setzt weiterführende Statistik-Kenntnisse voraus, insbesondere eine gründliche Vertrautheit mit dem Regressionsmodell (s. Kap. 10, Teil II).

# 5. Kapitel

# Dreidimensionale Tabellenanalyse: Drittvariablenkontrolle und Kausalmodelle

Beim Übergang von eindimensionalen (univariaten) zu zweidimensionalen (bivariaten) Verteilungen haben wir gesehen, wie sich die »bedingten« Häufigkeiten einer Variablen Y unterscheiden können je nach Wert oder Kategorie $x_i$ einer zweiten Merkmalsdimension X, die gleichzeitig realisiert ist. Wenn die bedingten Verteilungen $(y|x_i)$ der relativen Häufigkeiten nicht identisch sind, sprechen wir von einem »Zusammenhang« der beiden Variablen X und Y. In diesem Kapitel wollen wir darstellen, wie sich eine bivariate Verteilung verändern kann, wenn über die Ausprägungen $z_i$ einer dritten Variablen Z unterschiedliche Bedingungen vorgegeben werden. Mit der Einführung einer solchen »Kontrollvariablen« (auch »Test«- oder »Drittvariable« genannt) bewegen wir uns ein Stück weit in Richtung einer kausaltheoretischen Betrachtungsweise.

## 5.1 Ein einführendes Beispiel

Wir beginnen mit einem Beispiel, das wir der Habilitationsschrift von Heinrich Best über »Struktur und Handeln parlamentarischer Führungsgruppen in Deutschland und Frankreich 1848/49« (Köln 1986) entnehmen. Auf Grund einer (hier nicht zu erläuternden) Analyse des Abstimmungsverhaltens konnten den Abgeordneten der Frankfurter Nationalversammlung von 1848/49 unterschiedliche Positionen auf einer Ideologie-Skala zugewiesen werden, die durch die Pole »linke« vs. »rechte« Orientierung definiert war. Diese Skala haben wir für die Zwecke unseres Beispiels dichotomisiert: Abgeordnete, die eher »linke« Positionen unterstützten, erhalten hier den Wert »1«, Abgeordnete mit eher »rechten« Präferenzen den Wert »-1«. Indikatoren für eine Rechtsorientierung sind z. B.

- die Zustimmung zu einem Versammlungsverbot bei Gefahren für die öffentliche Sicherheit und Ordnung (Abstimmung v. 26. 9. 48)
- die Forderung nach Ausschluß der Empfänger von Armenunterstützung vom Wahlrecht (20. 2. 49)

Beispiele »linker« Forderungen sind dagegen:

- Die Grundrechte sollen unabhängig vom Wohnsitz gelten (6. 12. 48)
- Keine Staatssteuer ohne periodische Bewilligung durch das Parlament (13. 2. 49)

Wie sich die Voten bei einer langen Serie von Abstimmungen zu einer quasi - metrischen Links/Rechts - Skala zusammenfassen lassen, kann hier nicht erörtert werden. Dazu sind komplexere Verfahren nötig, mit denen wir uns in diesem Einführungsskript nicht beschäftigen. Im Rahmen der Überprüfung eines sozialisationstheoretischen Erklärungsmodells untersucht Best u. a., in welchem Ausmaß unterschiedliche politische Tätigkeiten vor 1848 das Abstimmungsverhalten der späteren Abgeordneten schon vorgeprägt haben. Diese politischen Erfahrungen können wir vereinfachend in einer Variablen (POLERF) zusammenfassen, die wir schon in Abschnitt 2.5 gebildet haben. Ihre Kategorien und Codeziffern seien hier noch einmal wiederholt:

| | |
|---|---|
| Vor 1848 ausschließlich in politischen Ämtern tätig | (1) |
| Ausschließlich Illegale Aktivitäten vor 1848 | (3) |
| In beiden Bereichen tätig gewesen (»inkonsistente« Erfahrungen) | (2) |
| In keinem der beiden Bereiche tätig gewesen | (0) |

Hiermit war die Erwartung verbunden, daß Abgeordnete der Gruppe 1, die schon vor 1848 politische Entscheidungspositionen innehatten, eher dem rechten Flügel, Abgeordnete der Gruppe 3 hingegen eher dem linken Flügel zugehören würden (Best 1986, S. 501). Für die Abgeordneten der Gruppe 2 mit »inkonsistenten« politischen Erfahrungen erwartete man eine Position zwischen den beiden anderen Gruppen. Diese Hypothese kann man in der Vermutung zusammenfassen, daß sich zwischen den Variablen »Politische Erfahrung vor 1848 (POLERF)« und der Links/Rechts-Orientierung (LRO) ein statistischer Zusammenhang zeigen werde. Es wurde aber auch damit gerechnet, daß der Einfluß der persönlichen politischen Erfahrungen auf das Abstimmungsverhalten durch spätere Rollenerfordernisse des Abgeordnetendaseins wie auch durch den politisch-territorialen Kontext modifiziert sein könnte, in dem sich der einzelne Abgeordnete bewegte. Im Hinblick auf diese Dimensionen werden wir später die Dritt - oder Testvariablen definieren. Zunächst präsentieren wir die zweidimensionale Kontingenztabelle der Variablen »Politische Erfahrung vor 1848« und »Links/Rechts-Orientierung« (s. *Abb. 5.1*).

Die fehlenden Werte (»Missing Observations«) ergeben sich daraus, daß nicht alle 809 Abgeordneten nach ihrer Links/Rechts-Einstellung klassifiziert werden konnten. Wir gehen davon aus, daß die Ergebnisse dadurch nicht verzerrt wurden. (Zum Problem der fehlenden Werte siehe Kap. 13, Teil II dieses Grundkurses.) Bevor wir mit der Analyse beginnen, sei noch einmal betont, daß es uns hierbei lediglich darum geht, die Grundelemente mehrdimensionaler Tabellenanalyse zu verdeutlichen. Der theoretischen und historischen Komplexität des Gegenstandes, den wir als Demonstrationsmaterial benutzen, werden wir dabei nicht gerecht. Dazu

Abb. 5.1: Zusammenhang von politischer Erfahrung und Links/Rechts-Orientierung

| | POLERF | | | | |
|---|---|---|---|---|---|
| COUNT / ROW PCT / COL PCT | Ik. polit. Erfahr. | nur Amts erfahrung | inkonsist. Erfah. | nur Opposition | ROW TOTAL |
| L-R-O | 0 | 1 | 2 | 3 | |
| -1 rechts | 249 | 83 | 40 | 59 | 431 |
| | 57.8 | 19.3 | 9.3 | 13.7 | 56.3 |
| | 60.1 | 56.5 | 50.6 | 46.8 | |
| 1 links | 165 | 64 | 39 | 67 | 335 |
| | 49.3 | 19.1 | 11.6 | 20.0 | 43.7 |
| | 39.9 | 43.5 | 49.4 | 53.2 | |
| COLUMN TOTAL | 414 | 147 | 79 | 126 | 766 |
| | 54.0 | 19.2 | 10.3 | 16.4 | 100.0 |

CRAMER'S V 0.10293

NUMBER OF MISSING OBSERVATIONS = 43

müßten auch komplexere statistische Verfahren eingesetzt werden. Der interessierte Leser sei auf die entsprechenden Kapitel in der Schrift von Best (1986) verwiesen.

Die Tabelle in *Abb. 5.1* bestätigt die Hypothese wenigstens der Tendenz nach. Die Abgeordneten der Gruppe 3 stellen mit 53.2 % einen um ca. 10 % höheren Anteil an Linksorientierten als die Gruppe 1; Gruppe 2 liegt mit 49.4 % dazwischen. Unter den »Unpolitischen« findet man mit knapp 40 % den geringsten Anteil an »linken« Abgeordneten. (Wir betrachten die Variable POLERF nicht als Ordinalskala, da zumindest den Unpolitischen a priori, vor der Zusammenhangsanalyse, keine eindeutige Rangposition im Verhältnis zu den drei anderen Gruppen zugewiesen werden konnte.)

Daß der Zusammenhang zwischen den beiden Variablen so schwach erscheint (siehe auch den niedrigen Wert für Cramers V) muß noch nicht bedeuten, daß vorgängige persönliche Erfahrungen das spätere Abstimmungsverhalten der Abgeordenten tatsächlich in so geringem Maße beeinflußt hätten. Ein Grund für die schwache Variablenbeziehung könnte in einer fehlerhaften »Messung« der Variablen zu finden sein. In unserem Beispiel mögen sowohl die Erfahrungskategorien als auch die LRO-Skala zu grob zusammengefaßt sein. Tatsächlich zeigt die Bestsche Detailanalyse mit der ursprünglich metrisch gebildeteten LRO-Skala, daß Abgeordnete,

die vor 1848 Funktionen in der lokalen Selbstverwaltung ausgeübt hatten, im Durchschnitt eher nach links votierten (ein arithmetisches Mittel über Null erreichten) als Abgeordnete, die Parlamentsmandate oder hohe Staatsämter innehatten oder in anderer Weise als Regierungs- und Standesvertreter tätig gewesen waren (und einen LRO-Wert unter Null erreichten). In unserer Erfahrungsvariablen sind diese beiden Abgeordnetengruppen jedoch zu einer einzigen Kategorie zusammengefaßt.

Wir wollen hier unser Augenmerk vor allem auf einen anderen Umstand richten, der die Schwäche des Zusammenhangs zweier Variablen u. U. zu erklären vermag: Die Wirkung eines Faktors (hier POLERF) auf eine abhängige Variable (hier LRO), kann selbst wiederum durch einen weiteren (»dritten«) Faktor (oder mehrere »dritte« Faktoren) beeinflußt sein. In unserem Beispiel könnten die regional unterschiedlichen Verfassungstraditionen, in deren Kontext die individuellen politischen Erfahrungen gemacht wurden, eine solche modifizierende Funktion ausgeübt haben. Best schreibt hierzu u. a.: »In Deutschland war die Verfassungstradition der Einzelstaaten das vermutlich wichtigste Differenzierungsmerkmal einer territorialen Segmentation politischer Erfahrungsmöglichkeiten. Nur dort, wo es Verfassungen und Volksvertretungen gab, war auch im unmittelbaren Umfeld der künftigen Abgeordneten der Nationalversammlung das Anschauungsmaterial für eine kompetitive Politik vorhanden, mit einer - wenn auch begrenzten - Konkurrenz zwischen Personen, Problemdefinitionen und Problemlösungen. Wir erwarten, daß ein solcher Wahrnehmungshintergrund auch ohne eigene Parlamentserfahrung Präferenzen für die Ausweitung und rechtliche Garantie von politischen Partizipationsmöglichkeiten begünstigte, also in unserem Untersuchungszusammenhang eine 'linke' Prädisposition setzte« (1986, S. 503). Hier wird zunächst einmal ein bivariater Zusammenhang zwischen der Variablen »Verfassungskontext bis 1848 (VK)« und der Variablen »LRO« vermutet. An anderer Stelle wird aber auch die weitergehende Hypothese geäußert, daß der Verfassungskontext die **Beziehung** zwischen den vorgängigen politischen Erfahrungen des Abgeordneten und seinem eher linken oder eher rechten Abstimmungsverhalten beeinflußt haben könnte. Man kann erwarten, daß in einer absoluten Monarchie die politischen Einstellungen von Amtsinhabern von den Einstellungen der »subversiv« Tätigen weiter entfernt waren und daß sie in dem so erfahrenen Gegensatz stärker und länger bei den Betroffenen nachwirkten als die Einstellungsdifferenzen, die in Verfassungsstaaten zwischen politischen Amtsträgern und oppositionellen Kräften auftraten. Beide Fragestellungen: a) diejenige nach dem direkten Einfluß der regionalen Verfassungstradition auf das Abstimmungsverhalten der Abgeordneten und b) diejenige nach dem Einfluß der Verfassungstradition auf das Ausmaß, mit der vorgängige politische Erfahrungen das spätere Abstimmungsverhalten beinflußt hat, sind analytisch

strikt voneinander zu unterscheiden. Prüfen wir zunächst die erste Hypothese:

Die Variable »Verfassungskontext vor 1848« enthält in unserer vereinfachten Version nur zwei Ausprägungen: (1) Absolute Monarchien, (2) Wahlregionen, die schon vor 1848 eine verfassungsstaatliche Ordnung erhielten.

Abb. 5.2: Zusammenhang von Verfassungskontext der Wahlregion vor 1848 und Links/Rechts-Orientierung

| | | Verfassungskontext | | |
|---|---|---|---|---|
| | COUNT I | | | |
| | ROW PCT I | absolute | Verfassung v. 18 | ROW |
| | COL PCT I | Mon. | | TOTAL |
| | I | 1 | 2 | |
| L-R-O | --------+ | --------+ | --------+ | |
| -1 | I | 305 | 126 | 431 |
| rechts | I | 70.8 | 29.2 | 56.3 |
| | I | 63.7 | 43.9 | |
| | +--------+ | --------+ | --------+ | |
| 1 | I | 174 | 161 | 335 |
| links | I | 51.9 | 48.1 | 43.7 |
| | I | 36.3 | 56.1 | |
| | +--------+ | --------+ | --------+ | |
| COLUMN | | 479 | 287 | 766 |
| TOTAL | | 62.5 | 37.5 | 100.0 |

PHI 0.19293

NUMBER OF MISSING OBSERVATIONS = 43

Wie die Tabelle in *Abb. 5.2* belegt, besteht ein deutlicher Zusammenhang zwischen dem Verfassungskontext und der Links/Rechts-Orientierung. 56.1 % der Abgeordneten aus Regionen mit verfassungsstaatlicher Ordnung, aber nur 36.3 % der Abgeordneten aus absoluten Monarchien tendieren nach »links«. Der Zusammenhang wäre vermutlich noch ausgeprägter, wenn nicht Bayern zwar schon vor 1830 eine Verfassungstradition entwickelt hätte, aber auch 1848 noch zum konservativen Kerngebiet gehörte. Eine detailliertere Analyse (siehe Best 1986) zeigt, daß Abgeordnete aus den »altkonstitutionellen« Einzelstaaten insgesamt weniger stark nach links tendierten als Abgeordnete aus Staaten, die erst während der 1830er Jahre Verfassungen erhalten hatten. Dies mag damit erklärbar sein, daß die politische Radikalisierung, die diese zweite Konstitutionalisierungswelle mit sich brachte, 1848/49 noch unmittelbar nachwirkte.

Die weitergehende Frage ist, ob sich die Wirkung vorgängiger persönlicher politischer Erfahrungen auf das Abstimmungsverhalten unterschiedlich gestaltete, je nachdem, in welchem Verfassungskontext sie sich vollzog. Diese Frage kann nur mit Hilfe einer dreidimensionalen Tabelle beantwortet werden: Die **Beziehung** zwischen LRO und POLERF wird jetzt getrennt untersucht - zum einen für Abgeordnete aus Wahlregionen mit absoluter Monarchie, zum anderen für Abgeordnete aus Wahlregionen mit verfassungsstaatlicher Ornung. Man sagt auch, die »Kontrollvariable« (hier: Verfassungskontext) werde »konstant gehalten«. Dieser Ausdruck ist vielleicht etwas mißverständlich, könnte er doch suggerieren, man wollte irgendwelche Daten versteckt manipulieren. Dem ist aber nicht so. Gemeint ist lediglich, daß die Beobachtungen in einer bestimmten Weise geordnet werden: Der Zusammenhang zwischen den beiden anderen Variablen, LRO und POLERF, wird nur für Teilgruppen untersucht, deren jeweilige Elemente sich hinsichtlich ihrer Werte $z_i$ (Kategorien) auf der dritten Merkmalsdimension nicht unterscheiden, »konstant« sind. Damit wird der mögliche Einfluß dieser dritten Variablen Z auf die Beziehung zwischen den beiden anderen Variablen X und Y innerhalb jeder Teilgruppe (innerhalb jeder »**Partialtabelle**«) ausgeschaltet. Etwaige Unterschiede **zwischen** den beiden Teilgruppen hinsichtlich des Zusammenhangs zwischen X und Y können dann aber gerade auf den Einfluß der Variablen Z zurückgeführt werden.

Die dreidimensionale Tabelle wird in SPSS$^x$ mit dem gleichen CROSSTABS-Kommando angefordert wie die zweidimensionale Tabelle (siehe Kap. 4). Die TABLES-Spezifikation wird lediglich mit einem weiteren BY um die Kontrollvariable erweitert.

```
CROSSTABS TABLES = LRO BY POLERF BY VK
STATISTIC          2
```

Für jede Kategorie der Kontrollvariable werden die Teiltabellen untereinander ausgedruckt. Wir haben sie hier nebeneinander gesetzt (siehe *Abb. 5.3*), um die Dreidimensionalität besser zu veranschaulichen.

Die Ergebnisse der dreidimensionalen Tabellenanalyse dürften ebenfalls durch Meßfehler im weitesten Sinne beeinträchtigt sein. Bei der VK-Variablen handelt es sich um eine territoriale Einteilung nach Wahlregionen. Wegen der Mobilität der Abgeordneten ist aber nicht sichergestellt, daß sie ihre Erfahrungen vor 1848 innerhalb des Verfassungskontextes machen konnten, der auch in ihrer Wahlregion gegeben war. Eine lückenlose Rekonstruktion der biographischen Mobilität ist jedoch nicht möglich.

Trotz dieser Probleme bestätigt die dreidimensionale Verteilung in *Abb. 5.3* die Hypothese: Der Zusammenhang zwischen POLERF und LRO ist unter den wechselnden Bedingungen des Verfassungskontextes unter-

Abb. 5.3: Dreidimensionale Verteilung von Links/Rechts-Orientierung, politischer Erfahrung und Verfassungskontext

| | | ABSOLUTE MON. POLERF | | | | VERFASSUNG V. 1848 POLERF | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Count / Row Pct / Col Pct | Ik. polit Erfahr | nur Amts erfahrun | INKONSIS T. ERFAH | NUR OPPO SITION | Ik. polit Erfahr | nur Amts erfahrun | INKONSIS T. ERFAH | NUR OPPO SITION |
| | | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 |
| L-R-O | | | | | | | | | |
| RECHTS | -1 | 200 | 41 | 21 | 43 | 49 | 42 | 19 | 16 |
| | | 65.6 | 13.4 | 6.9 | 14.1 | 38.9 | 33.3 | 15.1 | 12.7 |
| | | 63.9 | 73.2 | 80.8 | 51.2 | 48.5 | 46.2 | 35.8 | 38.1 |
| LINKS | 1 | 113 | 15 | 5 | 41 | 52 | 49 | 34 | 26 |
| | | 64.9 | 8.6 | 2.9 | 23.6 | 32.3 | 30.4 | 21.1 | 16.1 |
| | | 36.1 | 26.8 | 19.2 | 48.8 | 51.5 | 53.8 | 64.2 | 61.9 |
| | Column Total | 313 | 56 | 26 | 84 | 101 | 91 | 53 | 42 |
| | | 65.3 | 11.7 | 5.4 | 17.5 | 35.2 | 31.7 | 18.5 | 14.6 |
| | | CRAMER'S V .15260 | | | | CRAMER'S V .10276 | | | |

Abb. 5.4: Dreidimensionale Verteilung: Politische Erfahrung, Verfassungskontext vor 1848. Links/Rechts-Orientierung

Politische Erfahrung

| | | 0 | | 1 | | 2 | | 3 | |
|---|---|---|---|---|---|---|---|---|---|
| | COUNT / COL PCT | absolute Mon. | Verfassung v. 18 | absolute Mon. | Verfassung v. 18 | absolute Mon. | Verfassung v. 18 | absolute Mon. | Verfassung v. 18 |
| | | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| rechts | -1 | 200 | 49 | 41 | 42 | 21 | 19 | 43 | 16 |
| | | 63.9 | 48.5 | 73.2 | 46.2 | 80.8 | 35.8 | 51.2 | 38.1 |
| links | 1 | 113 | 52 | 15 | 49 | 5 | 34 | 41 | 26 |
| | | 36.1 | 51.5 | 26.8 | 53.8 | 19.2 | 64.2 | 48.8 | 61.9 |
| | COLUMN TOTAL | 313 | 101 | 56 | 91 | 26 | 53 | 84 | 42 |
| | | 75.6 | 24.4 | 38.1 | 61.9 | 32.9 | 67.1 | 66.7 | 33.3 |

schiedlich stark ausgeprägt. Von den Abgeordneten aus absoluten Monarchien, die vor 1848 politisch-oppositionell tätig gewesen waren (Gruppe 3), stimmten 48.8 % »links«, unter den vormaligen Amtsträgern (Gruppe 1) votierten nur 26.8 % in dieser Richtung. Die Abgeordneten mit in diesem Sinne inkonsistenter politischer Erfahrung tendierten mit einem überraschend starken Anteil (80.8 %) nach rechts. Ein anderes Bild erhält man in der zweiten Teiltabelle: Der Zusammenhang zwischen POLERF und LRO ist für Abgeordnete aus Wahlregionen mit verfassungsstaatlicher Ordnung deutlich schwächer. Insbesondere verringert sich der Unterschied zwischen den Abgeordneten mit Amtserfahrung und den Abgeordneten aus der politischen Opposition; die Prozentdifferenz (61.9 % - 53.8 %) ist von 22 auf 8 Punkte gesunken. Es bietet sich an, die Stärke dieses spezifizierenden Einflusses, den der Verfassungskontext auf die **Beziehung** zwischen POLERF und LRO ausübt, mit der Differenz (»2. Ordnung«) der Prozentdifferenzen (»1. Ordnung«) auszudrücken: $d^*\% = 22\% - 8\% = 14\%$. Der summarische Koeffizient, Cramers V, der alle Zellen der jeweiligen Teiltabelle berücksichtigt, ist hier weniger aussagekräftig.

Wenn sich der Zusammenhang zweier Variablen, Y und X, bei wechselnden Bedingungen, die mit den Ausprägungen $z_i$ der Kontrollvariablen Z gegeben sind, ändert, spricht man von einer **Interaktion** der Variablen X und Z in ihrer Wirkung auf Y: Richtung und/oder Stärke des Zusammenhangs zwischen X und Y, hängt davon ab, welcher Wert $z_i$ der Kontrollvariablen Z gleichzeitig realisiert ist.

Man kann in der dreidimensionalen Tabelle auch die Perspektive wechseln, indem man X- und Z-Variable in ihren Positionen tauscht. Dann läßt sich feststellen, daß der Einfluß des Verfassungskontextes auf das Abstimmungsverhalten durch die vorgängigen persönlichen Erfahrungen der Abgeordneten modifiziert wird. Das wird deutlich, wenn man in *Abb. 5.3* die Spaltenpaare, die untereinander verglichen werden, neu ordnet (s. oben *Abb. 5.4*):

Der Verfassungskontext der Wahlregion wirkte sich auf die Links/Rechts-Einstellungen der Abgeordneten relativ geringfügig aus, wenn diese über ausschließlich systemoppositionelle politische Erfahrungen verfügten; die Prozentdifferenz zwischen Abgeordneten aus absoluten Monarchien und denen aus Verfassungsstaaten beträgt in diesem Falle nur 13 Punkte. Stärker differenzierte der Verfassungskontext die politischen Orientierungen derjenigen Abgeordenten, die vor 1848 ausschließlich innerhalb der Amtshierarchie tätig gewesen waren; die Prozentdifferenz beträgt nun 27 Punkte. Wiederum zeigen diese unterschiedlichen Prozentdifferenzen die interaktive Wirkung der Variablen VK und POLERF auf LRO an: Auch hier beträgt die Prozentdifferenz zweiter Ordnung wieder

14 Punkte (27 % - 13 %). Die Aussage, Z spezifiziert den Einfluß von X auf Y, ist also »statistisch« identisch mit der Aussage: X spezifiziert den Einfluß von Z auf Y. Diese Identität führt zu der abstrakteren Formulierung: X und Z interagieren in ihrem Einfluß auf Y. Wie wir später noch sehen werden, bedeutet das nicht unbedingt, daß X und Z auch untereinander korrelieren.

Die dreidimensionale Kontingenztabelle enthält alle Informationen, die auch in den univariaten und bivariaten Häufigkeitsverteilungen der beteiligten Variablen zu finden sind, und ergänzt sie durch neue Einsichten: Die Veränderbarkeit oder Stabilität der bivariaten Verteilungen unter wechselnden Bedingungen, die mit den Kategorien der dritten Variablen gesetzt sind, wird erkennbar. So wie die univariaten Verteilungen als »Randverteilungen« der bivariaten Tabellen erscheinen, bilden die bivariaten Tabellen (hier POLERF mit LRO, VK mit LRO, VK mit POLERF) die **Marginalverteilungen** (»Marginaltabellen«) zur trivariaten Häufigkeitsverteilung. Und so, wie aus den bivaritaten Verteilungen (also den Zellenhäufigkeiten der zweidimensionalen Tabelle) die univariaten Verteilungen der beiden Variablen durch einfaches Summieren zu errechnen sind, lassen sich auch die bivariaten Verteilungen aus der dreidimensionalen Tabelle rekonstruieren. Die bivariate Verteilung von POLERF und VK (s. *Abb. 5.5*) zum Beispiel ergibt sich aus der dreidimensionalen Tabelle in *Abb. 5.3*.

Die Zellenbesetzungen der bivariaten Verteilung zwischen den beiden unabhängigen Variablen sind identisch mit den Spaltensummen der dreidimensionalen Tabelle. Aus den drei bivariaten Tabellen läßt sich aber nicht umgekehrt die dreidimensionale Tabelle konstruieren - ebensowenig, wie sich aus den univariaten Randverteilungen die Zellenhäufigkeiten der zweidimensionalen Tabelle erschließen lassen.

Die Kennzahlen, mit denen man die Stärke des statistischen Zusammenhangs zweier Variablen in den Partialtabellen ausdrückt, bezeichnet man als **bedingte** Korrelations- oder Assoziationskoeffizienten. So erhalten wir in unserem Beispiel neben den bedingten Prozentdifferenzen zwei bedingte V-Werte für den Zusammenhang zwischen POLERF (X) und LRO (Y): $V_{(y,x)|z(1)} = 0{,}15$ unter der Bedingung z(1)=: absolute Monarchie und $V_{(y,x)|z(2)} = 0{,}10$ unter der Bedingung z(2)=: verfassungsstaatliche Ordnung (s. *Abb. 5.3*). Für den bivariaten statistischen Zusammhang (»zero-order correlation«) war zuvor ein $V_{yx} = 0{,}10$ ermittelt worden (siehe *Abb. 5.1*).

Der Einfluß der dritten Variable Z auf die Beziehung zwischen zwei anderen Variablen X und Y kann also an zwei »Stellen« sichtbar wrden: einmal in der oder in den Differenzen zwischen nicht-konditionierten (bivariaten) und bedingten Assoziationskoeffizienten (einschließlich der

Abb. 5.5: Rekonstruktion der bivariaten Verteilung von polit. Erfahrung u. Verfassungskontext aus der dreidim. Tab. in Abb. 5.3

| | Politische Erfahrung | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | |
| Abs. Mon. | 313 | 56 | 26 | 84 | 479 |
| | 75.6 % | 38.1 % | 32.9 % | 66.7 % | 62.5 % |
| Verfassung | 101 | 91 | 53 | 42 | 287 |
| | 414 | 147 | 79 | 126 | 766 |

Prozentdifferenzen) und zum anderen in den Differenzen zwischen den bedingten Assoziationsmaßen. Wenn keine starken Differenzen zwischen den bedingten Koeffizienten bestehen, wenn vor allem keine Richtungsänderung des Zusammenhangs in den verschiedenen Teiltabellen zu beobachten ist, können die bedingten Koeffizienten zu einer neuen summarischen Maßzahl zusammengefaßt werden. Hierfür sind unterschiedliche Vorschläge gemacht worden, die sich darin unterscheiden, wie die bedingten Koeffizienten zu diesem Zweck jeweils gewichtet und gemittelt werden sollen. Die so gebildeten Kennzahlen bezeichnet man im Unterschied zu den »bedingten« als **partielle Assoziationsmaße**. Sie drücken den **spezifischen** Einfluß einer Variable X auf eine Variable Y aus, nachdem der Einfluß einer dritten Variable Z durch »Konstanthalten« ihrer Kategorien, wie oben gezeigt, neutralisiert worden ist. (Freilich bleibt offen, ob es nicht noch andere, unberücksichtigte Variablen gibt, die weiterhin die Beziehung zwischen X und Y beeinflussen.) In SPSS bzw. SPSS$^{x}$ ist (im Rahmen der Tabellenanalyse) nur für Gamma das partielle Assoziationsmaß (und auch dort nur im Integer-Mode des CROSSTABS-Kommandos) verfügbar. Definitionen einiger Partialkoeffizienten sind z. B. in Schmierer 1975, S. 113, 128 - 131 erläutert.

Fassen wir das bisher Gesagte in den wesentlichen Punkten zusammen: Ausgangspunkt unserer Betrachtung ist die Assoziation zweier Variabler,

X und Y, wie sie sich in einer zweidimensionalen Tabelle als positiv, negativ oder nicht vorhanden darstellt. Mit Hilfe einer dreidimensionalen Tabelle läßt sich überprüfen, ob diese Assoziation bestätigt, aufgehoben oder modifiziert wird, wenn der Einfluß einer dritten Variablen, Z, durch »Konstanthalten« ihrer Kategorien neutralisiert wird. »Konstanthalten« bedeutet: die Beziehung zwischen X und Y wird innerhalb von k , k = 1, 2, ... K, Teilgruppen untersucht, die durch die K Kategorien der Z-Variablen definiert sind. Innerhalb jeder Teilgruppe haben alle Fälle dieselbe Ausprägung auf der Merkmalsdimension Z; aber die Elemente unterschiedlicher Teilgruppen haben unterschiedliche Z-Ausprägungen. Die bivariate Verteilung von X und Y, die innerhalb jeder Teilgruppe beobachtet wird, ist somit eine **bedingte** Verteilung $(y,x \mid z_k)$: sie kommt unter der Bedingung zustande, daß alle **ihre** Fälle ein und dieselbe Z - Ausprägung aufweisen. Die in den bedingten Verteilungen (Partialtabellen) sichtbar werdende Assoziation von X und Y kann nicht mehr von Z beeinflußt sein, da Z innerhalb der Teilgruppen konstant ist; unterschiedliche X- und Y-Werte bei den einzelnen Fällen einer Teilgruppe können nicht durch unterschiedliche Z-Werte hervorgerufen worden sein. Wenn alle bedingten Assoziationen zwischen X und Y gleich sind und sich nicht von der unbedingten Assoziation unterscheiden, hat Z offensichtlich keinen Einfluß auf die Beziehung zwischen X und Y (es sei denn, er würde durch eine unbekannte bzw. nicht erfaßte weitere Variable verdeckt). Wenn die bedingten Assoziationskoeffizienten »stark« voneinander abweichen, beeinflussen sich X und Z wechselseitig in ihrer Beziehung zu Y. (Was »stark« in diesem Zusammenhang heißen soll, kann sowohl nach fallspezifischen Kriterien der theoretischen Relevanz als auch nach inferenzstatistischen Kriterien der »Überzufälligkeit« einer Differenz festgelegt werden; siehe Teil II, Kap. 8). Wir sprechen dann von einer **Interaktion** oder **Spezifikation**. Weichen die bedingten Assoziationskoeffizienten nicht stark voneinander ab, können sie zu **partiellen** Koeffizienten gemittelt werden. Die partiellen Assoziationskoeffizienten $r_{yx.z}$ indizieren die spezifische Einflußstärke einer Variablen X auf die Variable Y, wenn der Einfluß einer Drittvariablen Z durch Konstanthalten »ausgeschaltet« worden ist (»r« steht hier für irgendeinen der Assoziations- oder Korrelationskoeffizienten). Wechselt man die Position von X und Z in der dreidimensionalen Verteilung, erhält man mit $r_{yz.x}$ den spezifischen Einfluß von Z auf Y, wenn der Einfluß von X ausgeschaltet worden ist. Die beiden partiellen Assoziationsmaße können voneinander abweichen. (Wir erinnern uns: für die Ermittlung des Interaktionseffekts, der sich in der Differenz zweier bedingter Koeffizienten ausdrückt, ist es unbedeutend, welche der beiden Variablen als unabhängige Variable X und welche als Kotrollvariable Z eingesetzt wird.) In dem folgenden Abschnitt 5.2 werden wir neben der Interaktion noch andere typische Beziehungsmuster für drei Variablen er-

läutern, wie sie beim Übergang von der zweidimensionalen zur dreidimensionalen Kontingenztabelle sichtbar werden können.

## 5.2 Interpretationsschemata für Drei-Variablen-Modelle

In dem einführenden Beispiel des vorigen Abschnitts haben wir gezeigt, wie eine dritte Variable Z die zunächst beobachtete bivariate Verteilung zwischen zwei anderen Variablen, X und Y, verändern kann. Die dort beobachtete Veränderung war durch unterschiedliche Zusammenhangsstärken in den Teiltabellen $(y,x \mid z_k)$ gekennzeichnet. Dieses Muster haben wir als »Interaktion« oder »Spezifikation« bezeichnet. Es sind aber noch andere Beziehungsmuster möglich; die wichtigsten wollen wir in den folgenden Abschnitten vorstellen.

### 5.2.1 Interaktion und additive Multikausalität

Wir beginnen mit einem weiteren Beispiel zur Interaktion. Wir haben es ausgewählt, weil in ihm mit dem »Lebensalter« eine in den Sozialwissenschaften besonders »beliebte« Kontrollvariable auftritt, deren inhaltliche Interpretation aber oft unklar ist. Die Daten stammen aus einer Berliner Meinungsumfrage, die kurz vor der 1981er Wahl zum Abgeordnetenhaus durchgeführt wurde. Als abhängige Variable Y dient ein Index ALTEPOL, in dem Meinungen zur Wichtigkeit traditioneller politischer Ziele (wie ökonomischer Wohlstand, innere und äußere Sicherheit) relativ zu (damals) »neuen« politischen Zielvorstellungen (wie saubere Umwelt, Unterstützung alternativer Lebensformen) zusammengefaßt sind (siehe Thome 1985). Als unabhängige Variable X wird die formale Schulbildung betrachtet, für die hier nur zwei Ausprägungen definiert sind: kein Abitur ./. Abitur und mehr. Sie ist u. a. mit der Hypothese verknüpft, daß der mit höherer Schulbildung in der Regel verbundene höhere sozio-ökonomische Status und das (im Durchschnitt) höhere kognitive Niveau offener machen für neue, »postmaterialistische« Politik-Ziele, da die Befriedigug materieller Bedürfnisse in dieser Gruppe weitgehend gesichert ist. Die bivariate Tabelle (siehe *Abb. 5.6*) scheint diese Annahme zu bestätigen: Traditionelle Politik-Ziele werden von den Befragten mit hoher Schulbildung seltener für wichtig, häufiger für unwichtig gehalten als von den Befragten ohne Abitur. Durch Einführen der Kontrollvariable Alter wird dieses Bild erheblich modifiziert (siehe *Abb. 5.7*).

Unter der Kontrollbedingung »alt« verschwindet die Beziehung zwischen formaler Schulbildung und politischer Einstellung völlig; in der mittleren Altersgruppe ist sie nur schwach ausgeprägt, in der jüngsten

Abb. 5.6: Zusammenhang zwischen Schulbildung und Alte Politikpräferenzen (Berliner Umfrage 1981)

| | | Schulbildung | | |
|---|---|---|---|---|
| | | niedrig | hoch | |
| ALTEPOL | unwichtig | 87<br>11.3 % | 60<br>35.9 % | 147<br>15.7 % |
| | mittel | 370<br>48.2 % | 77<br>46.1 % | 447<br>47.8 % |
| | wichtig | 311<br>40.5 % | 30<br>18.0 % | 341<br>36.5 % |
| | | 768<br>100 % | 167<br>100 % | 935<br>100 % |

dagegen relativ stark. Es liegt also eindeutig eine Spezifikation (Interaktion) vor.

Unterschiede zwischen Altersgruppen können in mindestens zweierlei Weise theoretisch gedeutet werden: einmal als sog. Lebenszyklus-Effekt (mit zunehmendem Alter werden Menschen tendenziell konservativer) und zum anderen als sog. Generationen-Effekt: Die verschiedenen Jahrgangsgruppen (»Kohorten«) unseres Beispiels sind unter sehr unterschiedlichen gesellschaftlichen Bedingungen aufgewachsen: die älteste Gruppe vor dem Ende des 2. Weltkrieges, die mittlere während der Wiederaufbauphase bis Mitte der 60er Jahre (unter konservativer Regierung), die jüngste Gruppe in der folgenden Zeit hohen ökonomischen Wohlstands (und sozial-liberaler Regierungsdominanz). (Die Alterseinstufung erfolgte nicht nach dem Geburtsjahr der Befragten, sondern danach, in welche Zeitperiode die Lebensjahre zwischen 14 und 18 fielen.) Solche unterschiedlichen Erfahrungen während der Jugendzeit prägen, so lautet eine oft vertretene sozialisationstheoretische Hypothese, die gesamte weitere Lebens-

Abb. 5.7: Dreidimensionale Verteilung von Alter/Generation, Schulbildung und Politikpräferenz; Spezifikation der bivariaten Verteilung in Abb. 5.6

| | | Alter/Generation | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | "alt" | | "mittel" | | "jung" | | |
| | | Schulbildung | | | | | | |
| | | niedr. | hoch | niedr. | hoch | niedr. | hoch | |
| ALTEPOL | wichtig | 172 | 15 | 95 | 12 | 44 | 3 | 341 |
| | | 46.6% | 46.9% | 38.2% | 20.3% | 29.3% | 3.9% | 36.5% |
| | mittel | 185 | 16 | 120 | 30 | 65 | 31 | 447 |
| | | 50.1% | 50.0% | 48.2% | 50.8% | 43.3% | 40.8% | 47.8% |
| | unwichtig | 12 | 1 | 34 | 17 | 41 | 42 | 147 |
| | | 3.3% | 3.2% | 13.7% | 28.8% | 27.3% | 55.3% | 15.7% |
| | | 369 | 32 | 249 | 59 | 150 | 76 | 935 |
| | | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |

führung einschließlich der politischen Einstellungen (wenn auch gewisse Modifikationen möglich bleiben). Diese unterschiedlichen »Lebenserfahrungen«, hervorgerufen durch Diskontinuitäten in der gesellschaftlichen Entwicklung, konstituieren »Generationen«. Ihre Mitglieder unterscheiden sich, zum Beispiel in ihren politischen Präferenzen, nicht (nur), weil sie sich in unterschiedlichen Phasen ihres individuellen Lebenslaufs befinden, sondern weil sie in der gleichen Lebensphase (vor allem während der Jugendzeit) stark unterschiedliche Erfahrungen gemacht haben.

Wenn die Untersuchungsdaten nur zu einem einzigen Zeitpunkt erhoben worden sind, können Generationen- und Lebenszykluseffekte nicht »statistisch« voneinander getrennt werden. Die Situation verbessert sich erst, wenn wiederholte Datenerhebungen über viele Jahre vorliegen, mit denen der Entwicklungspfad von Angehörigen unterschiedlicher Geburtsjahrgänge über verschiedene Lebensalterstufen hinweg verfolgt werden

kann (»Kohortenanalyse«)[4]. In diesem Skript können wir nur auf die unterschiedlichen Möglichkeiten der Interpretation der Altersvariable aufmerksam machen, ohne die Problematik weiter zu vertiefen. Im folgenden werden wir die Altersvariable im Sinne von »Generationen« interpretieren.

Wie in jeder dreidimensionalen Tabelle sind auch hier wieder zwei Fragestellungen analytisch strikt voneinander zu trennen:

(1) Ist die Einstellung zu den traditionellen Politik-Zielen von der Generationenzugehörigkeit abhängig?
(2) Beeinflußt die Generationenzugehörigkeit die **Beziehung** zwischen formaler Schulbildung und politischer Einstellung? Eine Antwort auf Fragen dieses Typs ist anders und in der Regel theoretisch komplexer zu begründen als Antworten zu (1)

Die dreidimensionale Tabelle in *Abb. 5.7* gibt auf beide Fragen eine positive Antwort. Die (bedingten) Prozentdifferenzen in den drei Teiltabellen sind unterschiedlich groß, es liegt also eine Interaktion von Alter und formaler Schulbildung im Hinblick auf die Bewertung traditioneller politischer Ziele vor. Hält man visuell die Schulvariable konstant und variiert die Altersvariable, zeigt sich außerdem, daß in jeder der beiden Bildungsgruppen die traditionellen Politikziele um so eher für unwichtig gehalten werden, je niedriger das Alter ist. Dieser Alters- oder Generationeneffekt ist bei den formal höher Ausgebildeten stärker als bei denen, die kein Abitur haben. (Damit wird auch der Interaktionseffekt noch einmal bestätigt.) Theoretische Erklärungen hierfür, müßten die sich wandelnden sozialen Funktionen der Schulbildung in der deutschen Vor- und Nachkriegsgesellschaft berücksichtigen.

Auch in diesem Beispiel gibt es neben der Interaktion eine Korrelation der beiden unabhängigen Variablen: Die jüngere Kohorte hat einen höheren Abiturienten-Anteil als die ältere. Daß eine Interaktion von X und Z auch ohne Korrelation zwischen X und Z (in der bivariaten Marginaltabelle) beobachtet werden kann, soll anhand des folgenden konstruierten Zahlenbeispiels (*Abb. 5.8*) demonstriert werden (siehe Schmierer 1975, S. 122 f.)

Daß die bivariate Marginaltabelle mit Phi = 0 aus der dreidimensionalen Tabelle rekonstruiert wurde, ist, wie in Abschn. 5.1 gezeigt, daran

---

[4] In der Kohortenanalyse definiert man zusätzlich einen sog. Periodeneffekt. Er tritt auf, wenn sich auf Grund irgendwelcher Ereignisse die Variablenwerte (Häufigkeitsverteilungen) der einzelnen Jahrgangsgruppen **unisono** im Niveau verschieben. Einer der drei Effekte läßt sich stets als Linearkombination der beiden anderen Effekte darstellen, so daß eine Trennung nur mit Hilfe theoretischer Zusatzannahmen möglich ist.

erkennbar, daß die Spaltensummen der trivariaten Verteilung mit den Zellenbesetzungen der bivariaten Verteilung identisch sind.

Abb. 5.8: Interaktion zwischen X und Z ohne Korrelation zwischen X und Z

a)

| | $x_1$ | $x_2$ |
|---|---|---|
| $z_1$ | 49 | 13 |
| $z_2$ | 45 | 12 |
| | 94 | 25 |

$\varphi_{xz} = 0$

b)

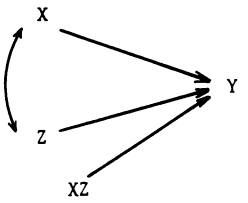
| | $z_1$ | | $z_2$ | |
|---|---|---|---|---|
| | $x_1$ | $x_2$ | $x_1$ | $x_2$ |
| $y_1$ | 12 | 7 | 11 | 10 |
| $y_2$ | 37 | 6 | 34 | 2 |
| | 49 | 13 | 45 | 12 |

$\varphi_{yx.z_1}=.26$ $\varphi_{yx.z_2}=.50$

Um die unterschiedlichen Strukturmuster von Variablenbeziehungen anschaulich darstellen zu können, führt man Pfeil- oder »Pfad«-Diagramme ein, wobei der Pfeil die vermutete Kausalrichtung angibt. Bisher haben wir folgende Beziehungsmuster kennengelernt:

(1) Interaktion von X und Z in ihrer Wirkung auf Y mit (bivariater) Korrelation zwischen X und Z (s. *Abb. 5.9*):

Abb. 5.9: Interaktion mit Korrelation zwischen unabh. Variable und Kontrollvariable




$r_{yx} \neq 0$

$r_{yz} \neq 0$

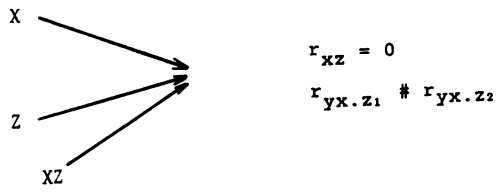
$r_{xz} \neq 0$

$r_{yx.z_1} \neq r_{yx.z_2}$

Die multiplikativ verknüpften X- und Z-Variablen (XZ) stehen für den Interaktionseffekt. Der gekrümmte Doppelpfeil deutet an, daß man die kausale Deutung der Korrelation $r_{xz}$ (zunächst) offen lassen will.

(2) Interaktion von X und Z in ihrer Wirkung auf Y ohne (bivariate) Korrelation zwischen X und Z (s. *Abb. 5.10*). So wie eine Interaktion

Abb. 5.10: Interaktion ohne Korrelation zwischen unabh. Variable und Kontrollvariable




ohne Korrelation zwischen X und Z auftreten kann, kann es auch eine Korrelation zwischen X und Z geben, ohne daß eine Interaktion zwischen X und Z vorliegt.

Ein Beispiel hierfür liefern die Tabellen in *Abb. 5.11, 5.12*). Die Rohdaten sind wiederum der schon erwähnten Studie von Best (1986) entnommen. Die abhängige Variable Y = Reichsidee wurde erneut aus dem Abstimmungsverhalten der Abgeordneten der Frankfurter Nationalversammlung konstruiert, diesmal aber im Hinblick auf die Bevorzugung großdeutscher oder kleindeutscher Lösungen des Verfassungskonflikts. Erste unabhängige Variable X ist eine territoriale Segmentation, bei der die einzelnen Wahlregionen wie folgt gruppiert wurden:

Gruppe 1: Oesterreich, Boehmen, Altbayern
Gruppe 2: Rheinpreußen, Franken, süddeutsche Klein- und Mittelstaaten
Gruppe 3: Altpreußen, Schlesien, Sachsen, mittel- und norddeutsche Klein- und Mittelstaaten

Die bivariate Tabelle (in *Abb. 5.11*) zeigt den erwarteten Zusammenhang. Die Abgeordneten der »nördlichen« Staaten präferieren mehrheitlich die kleindeutsche, die Abgeordneten aus Österreich usw. wollen mehrheitlich die großdeutsche Lösung.

Abb. 5.11: Zusammenhang zwischen Reichsidee und und territorialer Zugehörigkeit

| Count / Row Pct / Col Pct | ISÜDL.REG IION I | MITTELRE G. 1I | NÖRDL. R EGION 2I | 3I | Row Total |
|---|---|---|---|---|---|
| -1 großdeutsch | 170 | 109 | 102 | | 381 |
| | 44.6 | 28.6 | 26.8 | | 47.1 |
| | 74.2 | 54.0 | 27.0 | | |
| 1 kleindeutsch | 59 | 93 | 276 | | 428 |
| | 13.8 | 21.7 | 64.5 | | 52.9 |
| | 25.8 | 46.0 | 73.0 | | |
| Column Total | 229 | 202 | 378 | | 809 |
| | 28.3 | 25.0 | 46.7 | | 100.0 |

Cramers V .40529

Abb. 5.12: Dreidimensionale Verteilung von Reichsidee, territorialer Zugehörigkeit und persönlichem Religionsbekenntnis

| | Nicht-Katholiken | | | | Katholiken | | | |
|---|---|---|---|---|---|---|---|---|
| Count / Row Pct / Col Pct | ISÜDL.REG IION I | MITTELRE 1I | NÖRDL. R EGION 2I | 3II | ISÜDL.REG IION | MITTELRE G. 1I | NÖRDL. R EGION 2I | 3I |
| -1 großdeutsch | 11 | 51 | 74 | II | 156 | 56 | 15 | |
| | 8.1 | 37.5 | 54.4 | II | 68.7 | 24.7 | 6.6 | |
| | 57.9 | 47.7 | 23.2 | II | 76.1 | 61.5 | 36.6 | |
| 1 kleindeutsch | 8 | 56 | 245 | II | 49 | 35 | 26 | |
| | 2.6 | 18.1 | 79.3 | II | 44.5 | 31.8 | 23.6 | |
| | 42.1 | 52.3 | 76.8 | II | 23.9 | 38.5 | 63.4 | |
| Column | 19 | 107 | 319 | | 205 | 91 | 41 | |

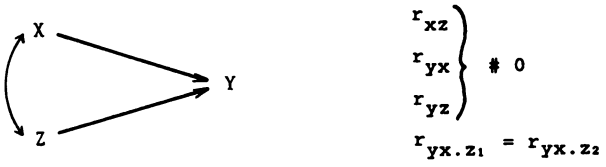
Cramers V .25785 Cramers V .27873

Sicherlich besteht auch 1848 noch ein starker Zusammenhang zwischen der regionalen und der konfessionellen Zugehörigkeit der Abgeordneten. Es entsteht also die Frage, ob das Votum für großdeutsche oder kleindeutsche Antworten eher durch Konfessions- als durch Staatsloyalität bestimmt war. Best (1986) schreibt hierzu: »Es zeichnet sich hier eine komplizierte Gemengelage konfessioneller und nichtkonfessioneller Determinanten ab, die sich erst klärt, wenn man die Ebene territorialer Aggregate verläßt und die individuellen Variablenzusammenhänge im Rahmen eines kausalanalytischen Ansatzes überprüft. Exemplarisch zugespitzt lautet die Frage, ob ein preußischer Abgeordneter kleindeutsch votierte, weil er Preuße oder weil er Protestant war. Die Antwort liegt bei den preußischen Abgeordneten katholischer Konfession« (S. 601). Wir führen demgemäß die Konfessionszugehörigkeit der Abgeordneten als Kontrollvariable ein (s. *Abb. 5.12*).

Da die Regionen der Gruppe 2 in beiden Teiltabellen dieselbe relative Position zwischen den beiden andereren Territorien einnehmen, beschränken wir uns auf den Vergleich der Gruppen 1 und 3. Die Prozentdifferenzen, 57.9 % - 23.2 % und 76.1 % - 36.6 %, sind in beiden Konfessionsgruppen nahezu gleich; eine nennenswerte Interaktion hat nicht stattgefunden. Das wird auch durch die beiden bedingten Assoziationskoeffizienten V = 0,26 und V = 0,28 belegt. Die territoriale Zugehörigkeit beeinflußt das groß- oder kleindeutsche Votum weitgehend unabhängig von dem persönlichen Religionsbekenntnis. Ebenso wirkt sich auch die Konfessionszugehörigkeit nahezu unabhängig von der territorialen Zugehörigkeit auf das Abstimmungsverhalten in dieser Frage aus. In jeder der drei territorialen Gruppierungen übersteigt der Prozentsatz der Protestanten für die kleindeutsche Lösung den Prozentanteil der Katholiken, die für die kleindeutsche Lösung stimmen, um 13 bis 18 Punkte. Wir wollen diese geringfügigen Schwankungen in den Prozentdifferenzen als unerheblich betrachten. Die beiden Variablen (konfessionelle und territoriale Zugehörigkeit) wirken also nicht interaktiv (»multiplikativ«), sondern »additiv« auf die Voten zu groß- versus kleindeutschen Konzeptionen. Graphisch läßt sich dieses Muster wie in *Abb. 5.13* darstellen.

Jeder der Abgeordneten, dessen Staatsloyalität in eine andere Richtung weist als seine konfessionelle Loyalität muß den Konflikt dieser beiden Einflußkomponenten in sich austragen. Das schließliche Resultat ist weitgehend unabhängig davon, ob der Konflikt zwischen nicht-katholischer Konfession und Zugehörigkeit zu einem katholisch dominierten Staat oder zwischen katholischer Konfession und Zugehörigkeit zu einem protestantisch dominierten Staat auszutragen ist. Der Anteil (42.1 %) an Nicht-Katholiken aus katholisch dominierten Regionen, die als Verfechter einer kleindeutschen Lösung auftreten, ist um ca. 35 Prozent niedriger als der entsprechende Anteil von Nichtkatholiken aus protestantisch domi-

Abb. 5.13: Additive Multikausalität mit Korrelation der beiden unabhängigen Variablen



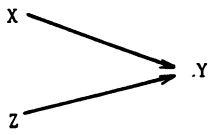

nierten Regionen. Andererseits votieren die Katholiken aus protestantisch dominierten Regionen nur zu 36.6 %, d. h. mit einem um 40 % niedrigeren Anteil für die großdeutsche Lösung als ihre Konfessionsbrüder aus katholisch dominierten Regionen.

Daß die bedingten Assoziationskoeffizienten in diesem Falle niedriger sind als die nicht-konditionierte Korrelation zwischen X und Y, ist durch die Korrelation zwischen X und Z verursacht. Die »Algebra« derartiger Beziehungsmuster, wird im Rahmen der sog. Pfadanalyse entwickelt, die zu den komplexeren Methoden gehört, die wir in diesem Skript nicht besprechen. Wenn die Kontrollvariable Z **nicht** mit der unabhängigen Variablen X korreliert (und auch keine Interaktion vorliegt), müssen die bedingten Assoziationen zwischen X und Y (unter Konstanthalten von Z) betragsmäßig größer sein als die Assoziation in der bivariaten Marginaltabelle (siehe *Abb. 5.14*).

### 5.2.2 Scheinkausalität

Dieses Beziehungsmuster wird in den meisten Statistik-Lehrbüchern mit folgendem Beispiel erläutert: Bei einer Erhebung regionaler Einheiten im Deutschland der Zwischenkriegszeit konnte man feststellen, daß ein starker Zusammenhang besteht zwischen der Anzahl der Störche, die in den einzelnen Regionen nisteten, und der Höhe der dort registrierten Geburtenziffer. Je geringer die Zahl der Störche, desto weniger Babies wurden geboren. Die Korrelation war zweifelsfrei vorhanden, obwohl eine Kausalbeziehung zwischen diesen beiden Variablen nach heutigem Wissen auszuschließen ist. Wie aber kann eine Korrelation zwischen zwei Variablen auftreten, ohne daß zwischen ihnen ein kausaler Zusammenhang besteht? Des Rätsels Lösung liegt in der Funktion der Drittvariablen »Indu-

Abb. 5.14: Additive Multikausalität ohne Korrelation der beiden unabhängigen Variablen




$$r_{xz} = 0$$
$$r_{yx} \neq 0$$
$$r_{yz} \neq 0$$
$$r_{yx.z_1} = r_{yx.z_2}$$
$$|r_{yx.z}| > |r_{yx}|$$

strialisierungsgrad«, die in der bivariaten Beziehung nicht berücksichtigt wurde und deshalb ihren Einfluß unkontrolliert geltend machen konnte: Ein höherer Industrialisierungsgrad mindert sowohl die Zahl der Störche als auch die Zahl der Babies. Wird der Industrialisierungsgrad als Kontrollvariable mit berücksichtigt, verschwinden in den Teiltabellen die Korrelationen zwischen Storchenzahl und Geburtenziffer. Wenn solche Phänomene auftreten, sprechen wir von »Scheinkausalität«. Gebräuchlicher ist der Ausdruck »Scheinkorrelation«. Er ist aber mißverständlich, da nicht die Korrelation »scheinbar« ist (sie besteht tatsächlich), sondern die daraus gefolgerte Kausalbeziehung. (Gelegentlich wird auch der Begriff »Unechte Korrelation« benutzt.)

Eine scheinkausale Beziehung kommt in der Forschungspraxis nur selten in reiner, »idealtypischer« Form vor, in der eine Korrelation nach Einführen einer Kontrollvariablen völlig verschwindet. In vielen Untersuchungen erhält man jedoch Variablenkonstellationen, die sich diesem Muster nähern. Da ihre »Logik« am deutlichsten in der idealtypischen Form sichtbar wird, wollen wir die Scheinkausalität anhand eines fiktiven Zahlenbeispiels erläutern. Dazu konstruieren wir eine andere Beziehung zwischen dem Abstimmungsverhalten (0 = großdeutsch/ 1 = kleindeutsch), der Konfessionszugehörigkeit (0 = katholisch/ 1 = nicht-katholisch) und der territorialen Zugehörigkeit (vergl. *Abb. 5.12*). Die bivariate Ausgangstabelle ist (bei veränderten Randverteilungen) in *Abb. 5.15 a* dargestellt.

Der Einfachheit wegen unterscheiden wir nur noch zwei Regionalgruppen: 0 = südliche, 1 = nördliche Staaten. Die trivariate Verteilung findet sich in *Abb. 5.15 b*.

Die in der bivariaten Tabelle enthaltene Assoziation zwischen Konfessionszugehörigkeit und Abstimmungsverhalten verschwindet in den bei-

Abb. 5.15: Fiktives Zahlenbeispiel zur Scheinkausalität (vergl.) Abb. 5.11 u. 5.12)

a) Bivariater Zusammenhang zwischen präferierter Reichsidee und Konfessionszugehörigkeit

| | kath. (0) | nicht-kath. (1) | |
|---|---|---|---|
| großdeutsch (0) | 240<br>55.6% | 128<br>38.1% | 368 |
| kleindeutsch (1) | 192 | 208 | 400 |
| | 432 | 336 | 768 |
| | $\varphi = 0.16$ | | |

b) Zusammenhang zwischen präferierter Reichsidee und Konfessinszugehörigkeit unter Kontrolle der territorialen Zuordnung

| | Südl. Reg.(0) kath. | nicht-kath. | Nördl. Reg.(1) kath. | nicht-kath. |
|---|---|---|---|---|
| großdeutsch | 224<br>66.7% | 96<br>66.7% | 16<br>16.7% | 32<br>16.7% |
| kleindeutsch | 112 | 48 | 80 | 160 |
| | 336 | 144 | 96 | 192 |
| | $\varphi_{yx \cdot z_0} = 0.0$ | | $\varphi_{yx \cdot z_1} = 0.0$ | |

den Partialtabellen der dreidimensionalen Verteilung. Sie wird durch den Einfluß der Regionalvariablen vollständig »erklärt«. Die Grundzüge dieser Numerik lassen sich (ohne formale Ableitung) wie folgt nachvollziehen:

1. Bei der hier verwendeten Kodierung (Zuordnung der Zahlen 0 und 1 zu den jeweiligen Ausprägungen) besteht eine »positive« Beziehung zwischen Region (Z) und Konfession (X): In den nördlichen Regionen ist der Anteil an Nicht-Katholiken höher als im Süden, ein hoher Regionalwert »geht tendenziell zusammen« mit einem hohen Konfessionswert. Das zeigt die Marginaltabelle, die wir aus der dreidimensionalen Tabelle in *Abb. 5.15 b* rekonstruieren (siehe *Abb. 5.16*)
2. Ebenso korreliert die Regionalvariable (Z) positiv mit dem Abstimmungsverhalten (Y). Abgeordnete aus den nördlichen Regionen votieren eher kleindeutsch (1) als süddeutsche Abgeordnete (siehe *Abb. 5.15 b*)

Diese Beziehungen lassen sich im Diagramm der *Abb. 5.17 a* veranschaulichen.

Um die Allgemeinheit dieses Beziehungsmusters auszudrücken, verwenden wir hier nicht das Phi-Symbol, sondern bezeichnen mit r ein beliebiges Korrelationsmaß, das positive und negative Werte annehmen kann. Die Deduktion mit $r^*$ gilt exakt nur für bestimmte Korrelationskoeffizienten, wie z. B. Phi in der 4-Felder-Tafel. Die Ableitung der Korrelationskoeffizienten aus dem Strukturmodell in *Abb. 5.17 a* kann man sich wie folgt plausibel machen: Wenn hohe (niedrige) Werte in Z hohe (niedrige) Werte in X **und** in Y nachsichziehen, bedeutet das, daß hohe (niedrige) Werte in X mit hohen (niedrigen) Werten in Y korrespondieren. Das aber heißt nichts anderes, als daß eine positive Korrelation zwsichen X und Y besteht, wenn der Einfluß von Z nicht (durch »Konstanthalten«) ausgeschaltet wird.

Übrigens änderte sich an der Sachlage nichts, würden wir die dichotomen Variablen so kodieren (die Nullen und Einsen für die Kategorien vertauschen), daß negative Beziehungen zwischen Z und X sowie zwischen Z und Y zustände kämen. Auch in diesem Falle wäre $r_{yx}$ positiv (siehe *Abb. 5.17 b*)

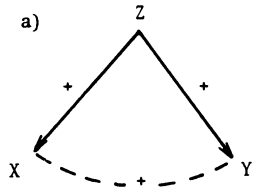
Möglich sind auch Muster, in denen die Kontrollvariable mit X (Y) positiv und mit Y (X) negativ korreliert. Diese umgekehrten Vorzeichen würden (wenn keine anderen Kausaleinflüsse wirksam wären) eine negative (»Schein«-)Korrelation zwischen X und Y implizieren.

Ausschlaggebend für die Interpretation einer Assoziation als »scheinkausal« sind nicht nur die numerischen Relationen zwischen bivariaten und bedingten bzw. partiellen Korrelationskoeffizienten, sondern die Stichhal-

Abb. 5.16: Fiktives Zahlenbeispiel, bivariater Zusammenhang zwischen territorialer Zuordnung und Konfessionszugehörigkeit

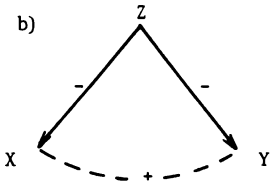
| | | Z: südl. | Z: nördl. |
|---|---|---|---|
| X | kath. | 336 | 96 |
| | | 70.0% | 33.3% |
| | nicht-kath. | 144 | 192 |
| | | 480 | 288 |

Abb. 5.17: Scheinkausalität




$r_{xz} > 0$
$r_{yz} > 0$
$r_{yx} > 0$

$r_{yx.z} = 0$
$r_{xz} > r_{yx} < r_{yz}$
$r^*_{yx} = r^*_{xz} \cdot r^*_{yz}$




$r_{xz} < 0$
$r_{yz} < 0$
$r_{yx} > 0$

$r_{yx.z} = 0$
$|r_{xz}| > |r_{yx}| < |r_{yz}|$

tigkeit, mit der die Kontrollvariable **theoretisch** als Faktor interpretiert werden kann, der den beiden anderen Variablen kausal **vorgeordnet** ist. In unserem fiktiven Beispiel nehmen wir also an, daß die territoriale Zugehörigkeit der Abgeordneten sowohl deren Konfession als auch deren Abstimmungsverhalten (mit) bestimmt.

Wir trennen somit analytisch das theoretische Konzept der **Kausalität** von dem empirisch-statistischen Konzept der **Korrelation**. Es ist aber nicht so, daß beide nichts miteinander zu tun hätten. Wir weisen nur darauf hin, daß bivariate Korrelationen ein sehr unzuverlässiger Indikator für das Vorliegen kausaler Beziehungen sind. Bessere Indikatoren erhalten wir mit den bedingten Koeffizienten, d. h. nach Einführung einer oder mehrerer Kontrollvariablen. Eine bestimmte Hypothese, ein **Modell** über kausale Beziehungen zwischen drei (oder mehr) Variablen, führt zu bestimmten **Deduktionen** hinsichtlich der bedingten und unbedingten Korrelationskoeffizienten. Indem diese Koeffizienten empirisch ermittelt werden, läßt sich das Modell testen. Stimmen die deduzierten Korrelationskoeffizienten nicht mit den empirischen überein, ist das Modell (oder ein Teil davon) widerlegt. Stimmen sie überein, ist es bestätigt, aber nicht endgültig bewiesen, denn man kann nie ausschließen, daß nicht-berücksichtigte Variablen das beobachtete Zusammenhangsmuster beeinflussen, daß es sich erneut ändert, wenn andere Variablen zusätzlich in das Modell aufgenommen werden.

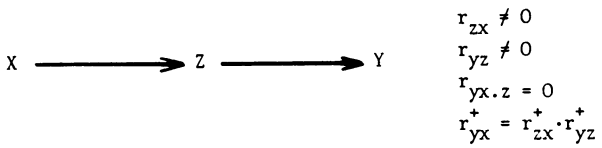
In der Praxis überlagern sich häufig scheinkausale und kausale Zusammenhänge zwischen X und Y. Dann bleibt auch nach Konstanthalten der Kontrollvariablen Z eine Korrelation zwischen X und Y erhalten, allerdings auf niedrigerem Niveau. (Dann müßte also ein durchgezogener Pfeil auch von X nach Y verlaufen, und es würde nicht mehr gelten: $r^*_{yx} = r^*_{xz} \cdot r^*_{yz}$.)

In dem folgenden Abschnitt stellen wir eine Beziehungsstruktur vor, die sich theoretisch von der Scheinkausalität erheblich unterscheidet, dennoch mit ihr oft verwechselt wird, weil beide Modelle zu den gleichen Deduktionen hinsichtlich der Korrelationskoeffizienten führen.

### 5.2.3 Intervention (Kausalkette)

Bei der Intervention wird die Kontrollvariable Z nicht wie im Modell der Scheinkausalität den beiden anderen Variablen kausal vorgeordnet, auch nicht (siehe additive Multikausalität) einer anderen unabhängigen Variablen kausal gleichgeordnet, sondern der unabhängigen Variablen X kausal nachgeordnet. Z nimmt also eine vermittelnde Position zwischen X und Y ein. Deshalb bezeichnet man dieses Modell auch als Kausalkette (siehe *Abb. 5.18*):

Abb. 5.18: Intervention

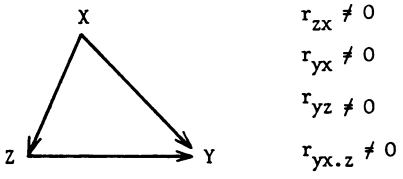



Ein Beispiel hierfür könnte die Beziehung zwischen Sozialschicht des Vaters (X), formaler Schulbildung des Sohnes (Z) und Einkommen des Sohnes (Y) sein. Die Hypothese wäre also: Je höher die Sozialschicht des Vaters, um so eher genießt der Sohn eine hohe Schulbildung, und infolgedessen erzielt er später auch ein hohes Einkommen. Wenn tatsächlich keine weiteren kausalen Einflußlinien auf Z und Y wirken, lassen sich aus diesem Modell die gleichen Korrelationskoeffizienten ableiten wie beim Modell der Scheinkausalität, insbesondere müssen die bedingten Assoziationskoeffizienten in den Partialtabellen gleich Null werden. Intuitiv kann man sich dies durch folgende Überlegung klarmachen: Wenn X, wie im Modell vorausgesetzt, nur über Z auf Y einwirkt, wird diese Einflußlinie »unterbrochen«, wenn man Z konstant hält. Dann varriert Y nur noch aufgrund von Einflüssen, die das Modell als »zufällig« behandelt und deshalb nicht expliziert.

Da sich aus diesem Modell die gleichen Assoziationskoeffizienten deduzieren lassen wie aus dem Modell der Scheinkausalität (vergl. *Abb. 5.18* mit *Abb. 5.17 a,b*) werden beide Modelle oft auch in ihrem theoretischen Gehalt fälschlich gleichgesetzt. So wollen Sozialwissenschaftler gelegentlich den Einfluß von Strukturvariablen X, wie »soziale Schicht«, auf Einstellungen oder Verhaltensmerkmale, Y, dadurch testen, daß sie abstraktere Werte-Indikatoren als Kontrollvariablen Z einsetzen. Wird daraufhin die partielle Assoziation $r_{yx.z} = 0 < |r_{yx}|$ beobachtet, sehen sich manche Interpreten veranlaßt, auch die kausale Relevanz von sozialer Schicht trotz $r_{zx} \neq 0$ und $r_{yz} \neq 0$ in der Nähe von Null anzusiedeln, da doch lediglich eine »Scheinkorrelation« zwischen ihr und der abhängigen Variablen bestehe. Eine solche Interpretation ist dann Unsinn, wenn die Wertevariable aufgrund einer bestimmten Theorie kausallogisch **zwischen** sozialer Schicht und Verhalten einzuordnen ist. Die partielle Assoziation $r_{yx.z} = 0$ **bestätigt** ja gerade die Hypothese einer Kausalkette, falls $r_{zx} \neq 0$ und $r_{yx} \neq 0$.

In der Realität wird die reine Kausalkette die strukturellen Beziehungen zwischen drei Variablen häufig nicht adäquat widerspiegeln. Realistischer ist in vielen Fällen das folgende Modell (siehe *Abb. 5.19*):

Abb. 5.19: Intervention plus direkter Effekt



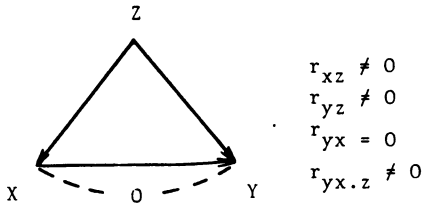

Die Variable X wirkt nicht nur »indirekt«, vermittelt über die »intervenierende« Variable Z auf Y, sondern auch »direkt« oder durch die Vermittlung anderer, nicht berücksichtigter Variablen. Dieses Modell ähnelt dem der additiven Multikausalität · mit dem Unterschied, daß nicht nur eine korrelative, sondern eine einseitig kausale Beziehung zwischen X und Z angenommen wird. In diesem Falle würde die Korrelation zwischen X und Y nicht durch Konstanthalten von Z verschwinden (sie könnte u. U. sogar größer werden, siehe den folgenden Abschnitt 5.2.4). Es wäre dann zu prüfen, ob es nicht eine andere intervenierende Variable $Z_2$ gibt, die eine zweite Wirkungskomponente von X nach Y vermittelt. Der sozioökonomische Status des Vaters könnte sich z. B. auch noch über soziale Kontaktfelder unabhängig von der formalen Schulbildung positiv auf den beruflichen und finanziellen Erfolg des Sohnes auswirken. Auch der »Pfad« zwischen Sozialschicht und Schulbildung könnte durch weitere Faktoren, z. B. die schichtspezifische Sprachkompetenz, vermittelt sein. Wenn immer wir fragen, **warum** wirkt eine Variable X auf eine Variable Y, fragen wir letztlich nach intervenierenden Variablen. Das gilt vor allem, wenn X eine jener »globalen« Strukturvariablen darstellt (wie »Industrialisierungsgrad«, »Soziale Schicht«), von denen sozialwissenschaftliche »Theorien« wimmeln. Je stärker wir einen zunächst »direkt« gemessenen Effekt in »indirekte« Effekte zerlegen können, desto vollständiger wird unser Wissen.

### 5.2.4 Suppression

Das Konzept der Suppression verdeutlichen wir zunächst wieder im idealtypischen Modell, das eine Umkehrung des Modells der Scheinkausalität darstellt: In der bivariaten Verteilung zwischen zwei Variablen X und Y wird kein oder nur ein relativ schwacher Zusammenhang festge-

stellt. Nach Einführung einer Kontrollvariablen Z, die zu den beiden anderen Variablen kausal vorrangig ist, wird in den Partialtabellen aber ein (stärkerer) Zusammenhang beobachtet (s. *Abb. 5.20*):

Abb. 5.20: Suppression




Ein Beispiel hierfür wäre folgende Variablenkonstellation: Y sei die Selbstmordrate in verschiedenen Regionen, X der Anteil der dort lebenden Juden, Z der jweilige Grad der Verstädterung. Gestützt auf die klassische Untersuchung Emile Durkheims über den Selbstmord, würde man einen negativen Zusammenhang zwischen X und Y vermuten (s. *Abb. 5.20 a*), gleichzeitig aber auch positive Zusammenhänge zwischen Z und X sowie zwischen Z und Y (s. *Abb. 5.20 b*):

Abb. 5.20 a

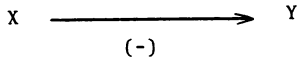
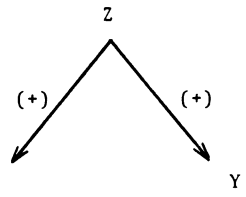



Abb. 5.20 b

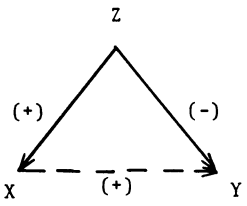



Gäbe es nur die beiden letztgenannten Kausalbeziehungen, nicht auch eine zwischen X und Y, so müßte sich eine positive bivariate Korrelation

zwischen X und Y beobachten lassen: Wenn Z ansteigt, nehmen die X- und Y-Werte gleichzeitig zu: $r^*_{yx} = r^*_{xz} \cdot r^*_{yz}$. Nun wirkt aber die tatsächlich bestehende **negative** Kausalbeziehung von X nach Y in Richtung einer negativen Korrelation zwischen X und Y: Wenn X ansteigt, fällt Y tendenziell ab. Die beiden Korrelationskomponenten, die auf gegenläufigen Kausallinien beruhen, heben sich, je nach Stärke der beiden Komponenten, tendenziell auf. Wenn man Z konstant hält, diesen kausalen Einfluß in den Partialtabellen somit ausschaltet, kann sich der von X ausgehende Einfluß auf Y sozusagen frei in dem Korrelationskoeffizienten entfalten.

Wenn zwischen X und Y nicht eine **negative** Kausalbeziehung (wie in dem Selbstmordbeispiel) besteht, sondern eine positive, so kann sie in der bivariaten Assoziationstabelle unterdrückt werden, wenn eine dritte Variable Z kausal vorrangig mit jeweils **umgekehrtem** Vorzeichen auf X und Y wirkt (s. *Abb. 5.21*).

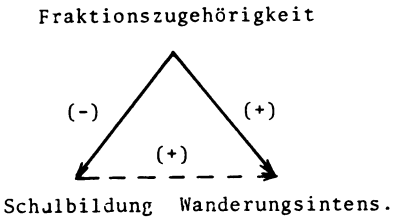
Abb. 5.21




Eine Spezialform der Suppression ist die sog. **Verzerrung**, bei der sich das Vorzeichen eines bivariaten Zusammenhangs nach Einführung einer Kontrollvariablen in den Partialtabellen umkehrt.

In der sozialwissenschaftlichen Forschungspraxis kommt die Suppression seltener in diesen idealtypischen Formen, häufiger vermischt mit Interaktionen vor. Einem entsprechenden Beispiel sind wir schon in Kap. 4, in den Abbildungen 4.14 a bis 4.14 c begegnet: Zwischen formaler Bildung (X) und Wanderungsintensität (Y) der Reichstagsabgeordneten haben wir zunächst nur einen schwachen Zusammenhang beobachtet. Nach Einführung der Kontrollvariablen (Z) »Fraktionszugehörigkeit«, wurde in den beiden Partialtabellen ein stärkerer Zusammenhang erkennbar; bei den SPD-Abgeordneten war er jedoch schwächer als bei den Abgeordneten anderer Parteien. Der dort beobachtete Suppressionseffekt läßt sich nun leicht anhand der *Abb. 5.22* erklären:

Abb. 5.22: Suppressionsbeispiel




Wenn die dichotomisierte Fraktionszugehörigkeit mit »1« für die SPD und mit »0« für andere Parteien kodiert wird, gibt es einen positiven Zusammenhang zwischen dieser Kontrollvariablen und der Wanderungsintensität (die SPD hat eher als andere Parteien ihre Abgeordneten in heimatferne Wahlkreise geschickt). Negativ ist hingegen der Zusammenhang zwischen Fraktionszugehörigkeit und formaler Bildung (die SPD hat einen überdurchschnittlich hohen Anteil an Abgeordneten ohne Abitur bzw. Hochschulstudium). Aus diesen beiden Kausalbeziehungen mit umgekehrtem Vorzeichen entsteht eine negative Korrelationskomponente für den Zusammenhang zwischen Schulbildung und Wanderungsintensität. (Die gleiche Situation wäre gegeben, wenn die Fraktionszugehörigkeit mit »0« für die SPD und »1« für die anderen Parteien kodiert worden wäre.) Diese negative Korrelationskomponente wird überlagert von einer positiven, die aus der Kausalbeziehung zwischen Schulbildung und Wanderungsintensität resultiert. Die beiden Korrelationskomponenten mit umgekehrtem Vorzeichen heben sich in der bivariaten Tabelle nicht völlig auf, führen aber zu einem betragsmäßig sehr niedrigen Assoziationskoeffizienten (siehe *Abb. 4.14 a*). Nach Konstanthalten der Fraktionszugehörigkeit, wenn also der von der Parteizugehörigkeit ausgehende kausale Einfluß ausgeschaltet ist, werden für die Partialtabellen höhere Assoziationskoeffizienten errechnet (*Abb. 4.14 b und c*).

Das Suppressionsmodell kann auch mit dem Interventionsmodell vermischt sein, wenn die Kontrollvariable (Z) nicht beiden anderen Variablen (X und Y) kausal vorgelagert ist, sondern eine vermittelnde Position zwischen ihnen einnimmt. Ein fiktives Zahlenbeispiel hierzu ist einem Auf-

satz von Kühnel/Terwey (1988) zu entnehmen. In der bivariaten Verteilung von Geschlecht (X) und Parteipräferenz (Y: CDU oder Nicht-CDU) zeigt sich keinerlei Zusammenhang (s. *Abb. 5.23*)

Nach Einführung der Kontrollvariable »Bildung« erhalten wir die Partialtabellen in *Abb. 5.24*.

Die in ihnen beobachten Prozentdifferenzen bestätigen die Erwartung, daß eine originäre Kausalbeziehung von Geschlecht zu Wahlabsicht verläuft. Nach der hier vorgenommenen Kodierung entsteht dadurch eine positive Korrelationskomponente für den Zusammenhang zwischen Geschlecht und Parteipräferenz (siehe *Abb. 5.25*).

Aus der negativen Kausalbeziehung zwischen Geschlecht und Bildung und der positiven zwischen Bildung und Parteipräferenz (für die CDU) resultiert aber (im Produkt) auch eine negative Korrelationskomponente für den (bivariaten) Zusammenhang zwischen X und Y. In dem fiktiven Zahlenbeispiel heben sich positive und negative Korrelationskomponenten auf.

Nicht alle Situationen, in denen eine partielle Korrelation betragsmäßig größer ist als die entsprechende bivariate Korrelation läßt auf einen Suppressionseffekt im hier erläuterten Sinne schließen. Betrachten wir zum Beispiel noch einmal das oben vorgestellte Modell der additiven Multikausalität ohne Korrelation zwischen X und Z (siehe *Abb. 5.14*). Wenn Z konstant gehalten wird, zeigt sich, daß die partielle Korrelation $|r_{yx.z}|$ größer ist als die nicht-konditionierte $|r_{yx}|$. Wenn Z mit X überhaupt nicht korreliert, kann diese Erscheinung nicht als Suppressionseffekt gedeutet werden. Warum $|r_{yx.z}| > |r_{yx}|$ sein kann, wird in Kap. 10 (Teil II) deutlich werden.

In diesem Zusammenhang soll noch auf ein weiteres, vielleicht überraschendes Phänomen hingewiesen werden: Wenn das Modell der additiven Multikausalität gemäß *Abb. 5.14* korrekt ist, gilt auch $|r_{xz.y}| > r_{xz} = 0$: Nach Konstanthalten der **abhängigen** Variablen wird in den Partialtabellen eine Korrelation zwischen X und Z beobachtet. Das heißt, es entsteht eine »unechte« partielle Korrelation, wenn eine kausal **nachrangige** Variable unsinnigerweise konstant gehalten wird (zur näheren Erläuterung siehe Davis 1971, S. 118 f.).

### 5.2.5 Abschließende Bemerkungen

Wir haben in diesem Abschnitt Kausalmodelle vorgestellt, die lediglich drei Variablen involvieren. Theorien so geringer Komplexität reichen aber kaum jemals aus, ein interessierendes Phänomen adäquat zu erklären. In der Regel muß eine größere Zahl von Variablen aufeinander bezogen werden, wobei oft auch die erklärenden Variablen untereinander in hierar-

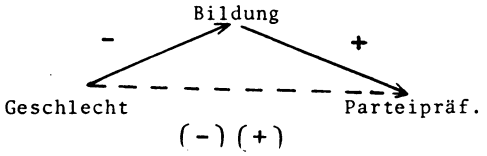
Abb. 5.23: Bivariate Verteilung von Geschlecht (1=weibl., 0=männl.) und Parteipräferenz (1=CDU, 0=andere) (fiktive Zahlen)

| | männl. | weibl. |
|---|---|---|
| CDU | 180 | 180 |
| andere | 220 | 220 |
| | 400 | 400 |

Abb. 5.24: Partialtabellen nach Einführung der Kontrollvariable "Bildung"

| | niedrige Bildung (0) | | hohe Bildung (1) | |
|---|---|---|---|---|
| | m | w | m | w |
| CDU | 30 | 75 | 150 | 105 |
| | 20 % | 30 % | 60 % | 70 % |
| andere | 120 | 175 | 100 | 45 |
| | 80 % | 70 % | 40 % | 30 % |
| | 150 | 250 | 250 | 150 |

Abb. 5.25: Suppression über Intervention



chischer Weise kausal verknüpft sind. Die explorative Konstruktion oder die empirische Überpüfung solcher Modelle erfordert komplexere Methoden der (multivariaten) Datenanalyse, die wir in diesem Grundkurs nicht behandeln können. Dabei bleiben jedoch die oben erläuterten Schemata kausaler Beziehungen durchaus anwendbar. Sie können Teilbeziehungen innerhalb eines komplexeren Modells adäquat erfassen; mehrere Variablen des Gesamtmodells bilden u. U. Bündel oder Blöcke, die als solche zueinander in Relationen stehen, wie wir sie für Drei-Variablen-Modelle definiert haben.

Der Sozialforscher wird nie in der Lage sein, die Werte einer abhängigen Variablen, die Häufigkeit mit der sie vorkommen, vollständig mit den Werten der unabhängigen Variablen zu erklären. Das bedeutet, es wird immer eine Reihe von Einflußfaktoren geben, die er in seinem Modell nicht spezifizieren kann. Unbekannte oder nicht erfaßbare Einflüsse bezeichnet man als »Zufallsfaktoren«. (Zu ihnen können auch Meßfehler gehören.) Der Forscher kann sie nicht einfach ignorieren, auch dann nicht, wenn er eine »vollständige« Erklärung gar nicht anstrebt. Das Problem liegt u. a. darin, daß er aus seinem Erklärungsmodell empirisch überprüfbare Kennzahlen (wie Korrelationskoeffizienten) nur dann korrekt ableiten kann, wenn er bestimmte (zutreffende) Annahmen über Eigenschaften dieser Zufallskomponenten macht. Zwei der wichtigsten Annahmen sind, daß sie sich »im Mittel« auf »lange Sicht« ausgleichen und daß die Zufallseinflüsse, die auf die abhängige Variable wirken, nicht mit Zufallseinflüssen korrelieren, die auf die unabhängigen Variablen wirken. Inhaltlich bedeutet diese Annahme, daß das »Modell« keine relevanten Erklärungsvariablen ausgelassen hat, die mit den berücksichtigten Erklärungsvariablen korrelieren. In Kap. 10 wird diese Problematik näher erörtert.

Diese knappen Bemerkungen sollten aber schon darauf hinweisen, daß wahrscheinlichkeitstheoretische und inferenzstatistische Konzepte nicht nur für die Stichprobentheorie, sondern auch für die Konstruktion von Erklärungsmodellen und deren empirischer Überprüfung wichtig sind.

Mehr Details und weiterführende Darstellungen zur Kausalanalyse mit Hilfe mehrdimensionaler Tabellen findet der Leser in Rosenberg 1968, Davis 1971, Hellevik 1984. Zur Diskussion über das Kausalitätskonzept siehe Falter/Ulbricht 1982. Sehr nützlich ist auch der Band von Hirschi/Selvin 1973, in dem weit verbreitete Analysefehler erörtert werden.

## 5.3 Ausblick auf die Analyse höherdimensionaler Tabellen (*)

Zur Darstellung vier- oder höherdimensionaler Verteilungen ist das mit dem CROSSTABS-Befehl in SPSS bzw. SPSS$^x$ produzierte Tabellenformat nicht geeignet. In SPSS$^x$ steht mit TABLES ein Unterprogramm zur Verfügung, das einen anderen Tabellenaufbau ermöglicht. Eine dieser neuen Varianten wollen wir hier etwas ausführlicher darstellen, da die entsprechende Programmbeschreibung in dem SPSS$^x$-Manual nicht so leicht zu lesen ist (allerdings gibt es ein umfangreiches Spezialmanual zu diesem Komplex).

Zunächst modifizieren wir in *Abb. 5.26* die dreidimensionale Tabelle aus *Abb. 5.3*. Die SPSS$^x$-Befehle hierzu lauten:

```
TABLES FORMAT = CWIDTH (20,5,20)
/FTOTAL = TOTAL 'SUMME'
/TABLE = LRO + TOTAL BY VK > POLERF + TOTAL
/STATISTICS = COUNT(LRO") CPCT (LRO":POLERF;VK)
/TITLE = '...'
```

Mit der FORMAT-Angabe regelt man die flächenmäßige Ausdehnung der Tabelle sowie andere graphische Details. Wir haben hier nur die Spaltenbreiten (CWIDTH) gegenüber den Default-Einstellungen geändert. Mit dem Kommando FTOTAL und dem zugeordneten Label 'SUMME' wird festgelegt, daß die Spalten und/oder Zeilensummen unter diesem Label mit ausgedruckt werden. Das TABLE - Kommando spezifiziert die Anordnung der Variablen. Als erstes wird die Zeilenvariable (in der Regel also die abhängige Variable, hier LRO) genannt. Wird die zuvor gebildete TOTAL-Variable mit einem Plus-Zeichen an sie angehängt, erscheinen die Spaltensummen als weitere Zeilengröße in der Tabelle. Nach dem ersten BY folgt, falls man die unabhängigen Variablen ineinander schachteln will, die in der Hierarchie höchstplazierte (Kontroll-)Variable, in unserem Beispiel also der Verfassungskontext VK. Die in sie »einzunistende« unabhängige Variabe POLERF wird mit dem Symbol »>« angefügt. Die wiederum mit dem Pluszeichen versehene TOTAL-Variable sorgt dafür, daß auch die Zeilensummen wie eine zusätzliche Spaltenvariable ausgedruckt werden. Man kann auch, wie bei CROSSTABS, Teiltabellen erstellen, indem man statt des Relationszeichens »>« ein weiteres BY einsetzt. Mit dem STATISTCS-Befehl werden jetzt nicht mehr (wie bei CROSSTABS) die verschiedenen Assoziationsmaße angefordert, sondern Informationen wie COUNT, die absoluten Fallzahlen der einzelnen Zellen, und CPCT, die Spaltenprozente. Mit der hinter COUNT in Klammern gesetzten Angabe LRO" wird festgelegt, daß die Fallzahlen innerhalb der Zellen ohne Label in der gleichen Zeile wie die Value-Labels der Zeilenvariable er-

Abb. 5.26: Abhängigkeit der Links/Rechts-Orientierung von politischer Erfahrung und Verfassungskontext vor 1848

| | Verfassungskontext | | | | | | | | Summe |
|---|---|---|---|---|---|---|---|---|---|
| | absolute Monarchie | | | | Verfassung vor 1848 | | | | |
| | Politische Erfahrungen | | | | Politische Erfahrungen | | | | |
| | keine politische Erfahrung | Amtsinhaber | Amtsinh. u subversiv | subversiv | keine politische Erfahrung | Amtsinhaber | Amtsinh. u subversiv | subversiv | |
| **Ideologie** | | | | | | | | | |
| rechts | 200 | 41 | 21 | 43 | 49 | 42 | 19 | 16 | 431 |
| | 63.9% | 73.2% | 80.8% | 51.2% | 48.5% | 46.2% | 35.8% | 38.1% | 56.3% |
| links | 113 | 15 | 5 | 41 | 52 | 49 | 34 | 26 | 335 |
| | 36.1% | 26.8% | 19.2% | 48.8% | 51.5% | 53.8% | 64.2% | 61.9% | 43.7% |
| | | | | | | | | | |
| **Summe** | 313 | 56 | 26 | 84 | 101 | 91 | 53 | 42 | 766 |
| | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

scheinen. Die Wiederholung dieser Angabe mit dem sog. Null-Label (") nach CPCT sorgt dafür, daß zwischen die Kategorien der Zeilenvariable LRO eine Leerzeile in die Tabelle eingefügt wird. Die weiteren Angaben nach dem Doppelpunkt in dieser Klammer besagen, daß die zu prozentuierenden Spalten durch die Variablen VK und POLERF definiert sind. Die Variable vor dem Doppelpunkt bezieht sich auf den Zähler, die Variable(n) nach dem Doppelpunkt auf den Nenner des Quotienten, der, mit 100 multipliziert, die Prozentzahl ergibt.

Das TABLES-Subprogramm bietet noch eine Reihe von Variationsmöglichkeiten für die Gestaltung der Tabellen. So ist es z. B. möglich, sämtliche unabhängige bzw. Kontrollvariablen als ineinandergeschachtelte Zeilenvariablen anzuordnen und die abhängige Variable als Spaltenvariable zu definieren. Diese Anordnungsform empfiehlt sich vor allem dann, wenn die Zahl der unabhängigen Variablen so groß ist, daß bei Spaltenschachtelung das Querformat einer Druckseite nicht genügend Platz für alle Spalten bietet.

Nach diesem Exkurs zur graphischen Gestaltung von Kontingenztabellen wenden wir uns nun wieder den Analysemethoden zu. Nachdem wir uns in diesem Kapitel relativ ausführlich mit dreidimensionalen Tabellen und den darauf bezogenen Kausalmodellen beschäftigt haben, wollen wir zum Schluß wenigstens noch andeuten, welche analytischen Möglichkeiten, aber auch, welche praktischen Probleme entstehen, wenn man mehr als drei Variablen gleichzeitig berücksichtigt. Zu diesem Zweck erweitern wir in unserem Analysebeispiel die bisherige dreidimensionale Verteilung um die Variable »Konfessionszugehörigkeit der Abgeordneten (KONF)« zu einer vierdimensionalen Tabelle. Da diese Variable einige fehlende Werte aufweist, reduziert sich die Fallzahl geringfügig. Erstellt wird die Tabelle in *Abb. 5.27* durch das eben besprochene SPSS$^x$- Subprogramm TABLES. Gegenüber dem obigen Beispiel mit der dreidimensionalen Tabelle, sind lediglich kleine Änderungen in den TABLE- und STATISTICS-Subkommandos nötig, die keiner weiteren Erklärung bedürfen:

```
/TABLE = LRO + TOTAL BY KONF > VK > POLERF + TOTAL
/STATISTICS = COUNT(LRO") CPCT (LRO":KONF,VK,POLERF)
```

Über die Einsichten hinaus, die die dreidimensionale Tabelle in *Abb. 5.26* vermittelt, macht die vierdimensionale Tabelle folgende Informationen **zusätzlich** verfügbar:

(a) Die univariate Häufigkeitsverteilung der konfessionellen Zugehörigkeit der Abgeordneten (hier nur in zwei Ausprägungen, katholisch ./. nicht-katholisch, kodiert). Dazu brauchen lediglich die entsprechenden Spaltensummen addiert zu werden.

Abb. 5.27: Abhängigkeit der Links/Rechts-Orientierung von politischer Erfahrung, Verfassungskontext und Konfessionszugehörigkeit

| | Konfession 1848 | | | | | | | | | | | | | | | | Summe |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | nicht kath. | | | | | | | | kath. | | | | | | | | |
| | Verfassungskontext | | | | | | | | Verfassungskontext | | | | | | | | |
| | absolute Monarchie | | | | Verfassung vor 1848 | | | | absolute Monarchie | | | | Verfassung vor 1848 | | | | |
| | Politische Erfahrungen | | | | Politische Erfahrungen | | | | Politische Erfahrungen | | | | Politische Erfahrungen | | | | |
| | keine Erfahrung | Amtsinhaber | Amtsinh. u. Irsiv | subv Irsiv | keine Erfahrung | Amtsinhaber | Amtsinh. u. Irsiv | subv Irsiv | keine Erfahrung | Amtsinhaber | Amtsinh. u. Irsiv | subv Irsiv | keine Erfahrung | Amtsinhaber | Amtsinh. u. Irsiv | subv Irsiv | |
| | In | | | | In | | | | In | | | | In | | | | |
| Ideologie | | | | | | | | | | | | | | | | | |
| rechts | 94 | 29 | 15 | 29 | 23 | 23 | 19 | 12 | 103 | 11 | 5 | 13 | 26 | 19 | | 4 | 425 |
| | 72.9% | 78.4% | 83.3% | 54.4% | 37.7% | 39.7% | 43.2% | 34.3% | 58.9% | 64.7% | 83.3% | 37.1% | 70.3% | 59.4% | | 66.7% | 57.2% |
| links | 35 | 8 | 3 | 16 | 38 | 35 | 25 | 23 | 72 | 6 | 1 | 22 | 11 | 13 | 8 | 2 | 318 |
| | 27.1% | 21.6% | 16.7% | 35.6% | 62.3% | 60.3% | 56.8% | 65.7% | 41.1% | 35.3% | 16.7% | 62.9% | 29.7% | 40.6% | 100% | 33.3% | 42.8% |
| Summe | 129 | 37 | 18 | 45 | 61 | 58 | 44 | 35 | 175 | 17 | 6 | 35 | 37 | 32 | 8 | 6 | 743 |
| | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

(b) Alle bivariaten Verteilungen, die die Konfessionsvariable mit jeder der anderen beteiligten Variablen bildet.

(c) Alle trivariaten Verteilungen, die die Konfessionsvariable mit jeweils zwei der anderen drei Variablen bildet. Sie wiederum enthalten die verschiedenen bivariaten Verteilungen unter wechselnden Bedingungen der jeweiligen Drittvariablen.

Zusätzlich zum Informationsgehalt der in ihr eingeschlossenen trivariaten Verteilungen beantwortet die vierdimensionale Verteilung noch zwei weitere Fragen, die analytisch zu trennen sind:

(d) Es kann der spezifische Einfluß der drei unabhängigen Variablen (POLERF, VK, KONF) auf die abhängige Variable (LRO) jeweils unter Konstanthalten der **beiden** anderen Variablen untersucht werden · zum Beispiel der Einfluß des Verfassungskontextes auf die Links/Rechts-Orientierung bei den Parlamentsabstimmungen unter Konstanthalten der vorgängigen politischen Erfahrungen **und** der Konfession der Abgeordneten.

(e) Es kann überprüft werden, ob die in der trivariaten Verteilung von LRO, POLERF und VK gefundene **Interaktion** zwischen POLERF und VK unter Konstanthalten der Konfessionsvariable stabil bleibt. Ebenso kann überprüft werden, ob auch Interaktionen zwischen zwei anderen Variablen, z. B. zwischen VK und KONF, bestehen und ob sie unter Konstanthalten der (dann) vierten Variable (POLERF) stabil bleiben.

Zunächst zum spezifischen Einfluß von VK auf LRO unter Konstanthalten von POLERF und KONF. Um die Interpretation einigermaßen übersichtlich zu halten, vergleichen wir hier nur die Erfahrungsgruppen 1 und 3, die Amtsinhaber und die »Subversiven«. Kombiniert mit den beiden Konfessionsgruppen ergeben sich somit 4 Bedingungen, die jeweils konstant zu halten sind:

(1) Nichtkatholische Amtsinhaber: Ist ihre Wahlregion eine absolute Monarchie, tendieren sie mit einem Anteil von 21.6 % nach »links«; kommen sie aus einem Verfassungsstaat, sind es 60.3 %. Das ergibt eine Prozentdifferenz von 38,7.

(2) Bei katholischen Amtsinhabern bewirkt der unterschiedliche Verfassungskontext eine LRO-Differenz von 40.6 % · 35.3 % = 5.3 %.

(3) Nichtkatholische »Subersive«: Derselbe Vergleich führt zu einer Prozentdifferenz von 65.7 % · 35.6 % = 30,1 %.

(4) Bei den katholischen Subversiven führt dieser Vergleich zu einem Vorzeichenwechsel: 33.3 % · 62.9 % = · 29.6 %. Allerdings ist hier die

Fallzahl sehr niedrig, nur 6 katholische »Subversive« kommen aus Verfassungsstaaten. Rechnet man die 8 Abgeordneten hinzu, die neben ihrer subversiven Tätigkeit auch schon Erfahrungen als Amtsinhaber aufweisen (und ebenfalls katholisch sind und aus Verfassungsstaaten kommen), so verändert sich diese Differenz auf 71.4 % - 62.9 % = 8.5 %.

Ein Einfluß des Verfassungskontextes auf die Links/Rechts-Orientierung im Abstimmungsverhalten bleibt also auch unter den erweiterten Kontrollbedingungen erhalten. Dieser Einfluß wird jedoch durch die Konfessionszugehörigkeit der Abgeordneten erheblich modifiziert. Am stärksten wirkt sich der Verfassungskontext auf das Abstimmungsverhalten der **nicht**-katholischen Amtsinhaber (d% = 38.7) aus, am schwächsten auf das Abstimmungsverhalten der katholischen Amtsinhaber (d% = 5.3). Für die Links/Rechts-Orientierung der katholischen Abgeordneten, sowohl der Amtsinhaber wie auch der »Subversiven«, spielt der Verfassungskontext kaum eine Rolle - im Gegensatz zu ihren nicht-katholischen Kollegen. Es gibt also eine Interaktion zwischen VK und KONF in ihrer Wirkung auf LRO. Sie wird durch die Differenz (2. Ordnung) der Prozentdifferenzen (1. Ordnung) ausgedrückt: Bei den Amtsinhabern beträgt sie 38,7 % - 5,3 % = 33,4 %, bei den »Subversiven« 30,1 % - 8,5 % = 21,6 %. Man kann nun auch noch eine Differenz 3. Ordnung bilden: 33,4 % - 21,6 % = 11,8 %. Sie zeigt an, daß die Interaktion (»1. Ordnung«) zwischen Verfassungskontext und Konfessionszugehörigkeit ihrerseits durch die politischen Erfahrungen der Abgeordneten modifiziert wird: Bei den Amtsinhabern ist sie etwas stärker als bei den Subversiven. Eine solche Verflechtung bezeichnet man als Interaktion 2. Ordnung. Sie läßt aber in unserem Falle weiterhin die Feststellung zu, daß die Konfessionszugehörigkeit den Einfluß des Verfassungskontextes auf die Links/Rechts-Orientierung der Abgeordneten weitgehend unabhängig von deren vorgängigen persönlichen Erfahrungen spezifiziert. Dabei stellen die katholischen Abgeordenten aus absoluten Monarchien anteilsmäßig mehr »Linke« (sowohl bei den Amtsinhabern, 35.3 %, als auch bei den »Subversiven«, 62.9 %) als ihre nichtkatholischen Kollegen (21.6 % und 35.6 %). Kommen sie hingegen aus Verfassungsstaaten, tendieren sie anteilsmäßig stärker nach »rechts« als ihre nicht-katholischen Kollegen.

Überprüfen wir nun noch ein früheres Ergebnis aus der dreidimensionalen Tabellenanalyse, wonach persönliche Erfahrungen und Verfassungskontext in ihrem Einfluß auf die Links/Rechts-Orientierung miteinander »interagieren« (angezeigt durch eine Differenz 2. Ordnung von 14 %). Die vierdimensionale Tabelle zeigt, daß bei konstant gehaltener Konfessionszugehörigkeit dieser Interaktionseffekt nur noch sehr schwach auftritt. Wir beschränken uns wiederum auf Amtsinhaber und »Subversive«. Bei

den **Nichtkatholiken aus Verfassungsstaaten** unterscheiden sich diese beiden Erfahrungsgruppen nur geringfügig in ihrem Anteil an Linksorientierten (65.7 % - 60.3 % = 5.4 %). Bei den **Nichtkatholiken aus absoluten Monarchien** ist diese Differenz, wie gehabt, größer: 35.6 % - 21.6 % = 14 %. Der (schwache) Interaktionseffekt 1. Ordnung zwischen VK und POLERF im Hinblick auf LRO bei den nicht-katholischen Abgeordneten ist durch die Differenz der Prozentifferenzen (Prozentdifferenz »2. Ordnung«), 14 % - 5.4 % = 8.6 % indiziert. Bei den **katholischen** Abgeordneten aus **absoluten Monarchien** ist die Differenz (1. Ordnung) der Linksanteile zwischen den beiden Erfahrungsgruppen größer als bei den Nicht-Katholiken: 62.9 % - 35.3 %. Bei den **Katholiken aus Verfassungsstaaten** ist der entsprechende Vergleich wiederum durch die geringe Fallzahl der Subversiven beeinträchtigt. Fassen wir sie mit den »Inkonsistenten« zu einer Gruppe zusammen (siehe oben) ergibt sich eine Differenz von 71.4 - 40.6 = 30,8 %, also eine ähnlich hohe Differenz wie bei den katholischen Abgeordneten aus absoluten Monarchien. Die Prozentdifferenz 2. Ordnung ist also nahe Null. Das heißt, innerhalb der Katholiken gibt es keinen nennenswerten Interaktionseffekt zwischen Verfassungskontext und politischer Erfahrung. Die politische Erfahrung hat bei ihnen in beiden Verfassungskontexten einen etwa gleich starken Einfluß auf die LRO; er ist deutlich höher als der Einfluß der persönlichen Erfahrungen bei den Nicht-Katholiken (bei denen, wie wir sahen, der Verfassungskontext stärker auf die ideologische Orientierung einwirkte). Der Interaktionseffekt zwischen Politischer Erfahrung und Verfassungskontext ist also bei Konstanthalten der Konfessionsvariable kaum noch vorhanden, während die Interaktion zwischen VK und KONF sowie POLERF und KONF in der vierdimensionalen Tabelle neu auftritt.

Die Interpretation vier- und mehrdimensionaler Tabellen mit der konventionellen Methode der Prozentvergleiche ist, wie das Beispiel zeigt, ziemlich mühsam. Leicht verfängt man sich im unübersichtlichen Gestrüpp immer neuer Kombinationen von Spaltendifferenzen, Differenzen von Differenzen usw. Die verschiedenen Haupt- und Interaktionseffekte unterschiedlicher Ordnung sind kaum auseinander zu halten. Um die vielfältigen Zusammenhangslinien besser voneinander trennen zu können, benötigt man stärker formalisierte Verfahren, wie **Log-lineare** bzw. **Logit-Modelle**, die aber mathematisch zu anspruchsvoll sind, um in diesem Skript vorgestellt zu werden. Im Kern geht es um die Anwendung des Regressionsmodells auf die Analyse von Kontingenztabellen. Kap. 12 in Teil II dürfte den Zugang zur entsprechenden Einführungsliteratur (siehe Aldrich/Nelson 1986; Jarausch/Arminger/Thaller 1985; Sensch 1987) erleichtern.

Zum Schluß noch eine Bemerkung, die an das oben aufgetretene Problem geringer Fallzahlen in einigen Spalten der vierdimensionalen Tabelle anknüpft. Es gibt z. B. nur 6 Abgeordnete, die katholisch sind, aus Wahlregionen mit verfassungsstaatlicher Tradition stammen und subversive politische Erfahrungen gemacht haben. Von denen haben nur zwei, rechnerisch also 33,3 %, bei Abstimmungen in der Frankfurter Nationalversammlung tendenziell »links« votiert. Andererseits haben 40,6 % der katholischen Abgeordneten, die **nicht** subversiv, sondern als Amtsinhaber tätig gewesen, aber ebenfalls in Regionen mit Verfassungstradition gewählt worden sind, »links« votiert. Erwartet hatte man jedoch eine Prozentdifferenz in umgekehrter Richtung. Was läßt einen Historiker oder Sozialforscher zögern, ein solches Ergebnis als Tatsache zu akzeptieren und seine Theorie als widerlegt oder ergänzungsbedürftig zu betrachten?

Handelte es sich bei seiner Untersuchungsgruppe nur um eine Stichprobe, eine »zufällige« Auswahl aus einer viel größeren (»Grund-)Gesamtheit von Abgeordneten, über die er letztlich Aussagen machen möchte, wäre die Antwort ziemlich klar: Von den 6 ausgewählten Personen mit dieser Merkmalskombination könnte man kaum annehmen, sie seien für die Gesamtheit aller Abgeordneten gleicher Merkmalskombination »repräsentativ«. Jedenfalls wäre man bei einer Auswahl von, sagen wir, 50 oder 100 Personen in dieser Gruppe eher bereit anzunehmen, daß ihre »Durchschnittsmeinung« mit der Durchschnittsmeinung der größeren Gesamtheit (nahezu) deckungsgleich wäre. (In den Kapiteln 6 bis 8, Teil II werden wir die entsprechenden Überlegungen genauer darstellen.) In unserem Analysebeispiel haben wir es jedoch (wenn wir von den ungeplant fehlenden Werten einmal absehen) mit der Gesamtheit der Fälle zu tun, über die wir Aussagen machen möchten: mehr Abgeordnete mit dieser Merkmalskombination gab es nicht. So ist der Befund: »Von ihnen votierten 33.3 %, von den vormaligen Amtsinhabern votierten (bei sonst gleicher Merkmalskombination) 40.6 % links« als deskriptive Aussage über die Grundgesamtheit gültig (wenn wir vom Problem der Meßfehler absehen). Er steht im Widerspruch zu der sozialisationstheoretischen Hypothese, wonach das Abstimmungsverhalten (die Links/Rechts-Orientierung) der Abgeordneten unbeschadet sonstiger Einflußfaktoren (wie Verfassungskontext und damit verbundene regionale Interessenlagen) von vorgängigen persönlichen Erfahrungen mit bestimmt sei. Aus dieser Hypothese war für die beiden hier betrachteten Gruppen eine Prozentdifferenz mit umgekehrtem Vorzeichen prognostiziert worden. (»Prognose« hier nicht als »Prädiktion«, sondern als »Postdiktion« verstanden).

Ein erstes Argument, das dennoch für die Beibehaltung der Theorie spricht, könnte darauf verweisen, daß die oben präsentierten Kreuztabellen auch eine Reihe von Daten (Prozentdifferenzen) enthalten, die die Theorie bestätigen. Zudem ließen sich unabhängig von diesem Datensatz

Ergebnisse aus anderen Untersuchungen zitieren, die die fragliche Theorie stützen. Es ist nicht sinnvoll, eine Theorie $T_1$ auf Grund eines einzigen Gegenbeispiels zu verwerfen, solange nicht eine alternative Theorie $T_2$ gefunden ist, die all das erklärt, was $T_1$ bisher zu erklären schien, und außerdem mit der Beobachtung vereinbar ist, an der $T_1$ scheiterte.

Weitere Argumentationsstrategien ergeben sich aus folgenden Überlegungen: Um alternative Theorien an einem Datensatz »testen« zu können, müssen sie in Modelle übersetzt werden, die aus den Hypothesen »erwartete« Beobachtungen, statistische Kenngrößen ableiten, die mit den empirischen Beobachtungen verglichen werden können. So haben wir z. B. in Kap. 4 aus der Hypothese der Unabhängigkeit zwischen zwei Variablen Erwartungswerte für die einzelnen Zellen einer Kontingenztabelle abgeleitet und sie mit den tatsächlich beobachteten verglichen. Die Vergleichsergebnisse wurden in der Kennzahl »Chi-Quadrat« zusammengefaßt, aus der dann weitere Maßzahlen abgeleitet wurden. In diesem Kapitel haben wir einfache Kausalmodelle erläutert, die Deduktionen über bedingte (bzw. partielle) und nicht-konditionierte Korrelationskoeffizienten erlaubten. Auch komplexere Modelle, die eine Vielzahl von Variablen umfassen, lassen sich überprüfen, indem man die aus ihnen gewonnenen »Erwartungswerte« verschiedener Art mit den entsprechenden Beobachtungen vergleicht. Man spricht in diesem Zusammenhang von der »Anpassungsgüte«, dem »Fit« eines Modells, der sich in bestimmten Kennzahlen (wie Chi-Quadrat) ausdrückt. Bei der Beurteilung, ob ein gegebenes Modell hinreichend gut oder besser als ein Alternativmodell an die beobachteten Werte »angepaßt« ist, können inferenzstatistische Kriterien (siehe Teil II) zu Rate gezogen werden. Mit ihrer Hilfe läßt sich, sehr grob gesagt, feststellen, wieweit die Differenz zwischen erwarteten und beobachteten Ergebnissen oder die Differenz zwischen der Anpassungsgüte des Modells A und der Anpassungsgüte des Modells B den Zufallseinflüssen zuzuschreiben ist, die bei einer Stichprobenziehung wirksam werden. Aber auch dann, wenn die Daten nicht aus einer Stichprobe, nicht aus einer ausgewählten Teilmenge stammen, sondern in der Grundgesamtheit erhoben wurden, empfiehlt sich häufig eine Modellevaluation auch nach inferenzstatistischen Kriterien. Dies vor allem dann, wenn die Grundgesamtheit (»Population«) selbst oder wenn Teilpopulationen, über die man spezielle Aussagen machen will, nur wenige Fälle umfassen - wie in dem oben zitierten Analysebeispiel mit Abgeordneten der Frankfurter Nationalversammlung. Wenn nur wenige Fälle vorliegen, kann man nicht damit rechnen, daß sich die im Modell nicht explizierten Zufallseinflüsse (siehe die Schlußbemerkung in Abschn. 5.2.5) wechselseitig aufheben. Vielmehr muß man davon ausgehen, daß sie die systematischen Einflüsse, die die Theorie nennt, so stark überlagern, daß die beobachteten Werte auch im Mittel relativ weit von den erwarteten abweichen, selbst wenn die Theorie

»korrekt« ist. Dieses Argument impliziert aber eine Konzeption, die jede **empirische** Population als Stichprobe aus einem **hypothetischen** Universum auffaßt; andernfalls müßten bei korrekter Modellspezifikation beobachtete und theoretisch erwartete Größen in der Population gleich sein. Die Hilfskonstruktion einer hypothetischen (Super-)Population läßt sich mit der wissenschaftslogischen These begründen, jede kausaltheoretische Erklärung sei in dem Sinne allgemein, daß sie keinerlei räumlich-zeitliche Beschränkungen ihrer Gültigkeit vorsehe. Auch wenn das zu erklärende Phänomen historisch einmalig sein sollte, impliziert die erklärende Theorie die Behauptung, daß jenes Phänomen wieder aufträte, wenn die in der Theorie spezifizierten Bedingungen erfüllt wären. Die Theorie besteht ja in der Relationierung von Wenn-Dann-Aussagen (wenn $x_1,...,x_k$, dann $y_1,...,y_m$). **Ob** die in dem Wenn-Teil formulierten Bedingungen überhaupt auftreten, kann sie offen lassen (oder mit Hilfe zusätzlicher Theorien zu prognostizieren versuchen). Die Anwendung der Inferenzstatistik bleibt allerdings bei dieser Konstruktion in dem Sinne »heuristisch«, als der Auswahlmechanismus nicht bekannt ist, die Anwendung der Inferenzstatistik aber einen Zufallsmechanismus bei der Auswahl voraussetzt.

Mit diesen Überlegungen soll übrigens nicht behauptet werden, theoretisch-allgemeine Erklärungen im Sinne der analytischen Wissenschaftstheorie seien das einzige legitime oder das vorrangige Arbeitsziel des Historikers. Wenn jedoch erklärende Theorien angestrebt werden, sollte über deren logische Struktur und über die methodologischen Implikationen Karheit bestehen. Und es sollte auch deutlich geworden sein, daß ein empirisch arbeitender Historiker elementare Kenntnisse der Wahrscheinlichkeitstheorie und Inferenzstatistik benötigt.

## Anhang:

## Das Rechnen mit Summenzeichen

### 1. Definition

In der Statistik werden häufig Ausprägungen ($x_i$) von Variablen (X), gelegentlich auch Konstanten (c) addiert. Will man z. B. das arithmetische Mittel $\bar{x}$ für eine Menge n von Untersuchungseinheiten berechnen, müssen die Merkmalsausprägungen ($x_1, x_2, \ldots, x_n$) aller Einheiten addiert (und anschließend durch n dividiert) werden. Es liegt nahe, für solche Additionsvorschriften ein »Kürzel« zu verwenden, konventionellerweise das griechische Sigma, $\Sigma$. Sind z. B. die X-Werte von 4 Untersuchungseinheiten zu addieren, schreiben wir

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = \sum_{i=1}^{n} x_i \quad , \; n = 4$$

Den Ausdruck auf der linken Seite des Gleichheitszeichens liest man als »Summe aller $x_i$-Werte für i = 1 bis 4«. Der sog. Laufindex i kann durch beliebige andere Buchstaben (z. B. j, k, m) ersetzt werden. Die Zahlen unterhalb und oberhalb des Summenzeichens geben jeweils den ersten und den letzten Wert an, der in die Summenbildung aufzunehmen ist. Häufig geht aus dem Kontext klar hervor, welche Werte zu summieren sind; dann kann man die Schreibweise weiter verkürzen:

$$\sum_{i=1}^{n} x_i = \sum_{i} x_i = \sum x_i$$

Die oben gegebene Definition schließt das Summieren von Produkten mit einfachem Laufindex ein:

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$$

### 2. Rechenregeln

Häufig sind Werte zu addieren, die mit einer Konstanten $c \neq 1, c \neq 0$ zu gewichten sind.

**Regel 1**

besagt, daß ein konstanter Faktor vor das Summenzeichen gezogen werden kann:

$$\sum_{i=1}^{n} cx_i = c \cdot \sum_{i=1}^{n} x_i$$

Diese Regel ergibt sich unmittelbar aus der Definition:

$$\sum cx_i = (cx_1 + cx_2 + \ldots + cx_n) = c(x_1 + x_2 + \ldots + x_n)$$

Falls die Konstante kein Faktor, sondern ein Summand ist, gilt

**Regel 2:**

$$(x_1 + c) + (x_2 + c) + \ldots + (x_n + c) = \sum x_i + n \cdot c$$

**Regel 3:**

$$\begin{aligned} \sum_{i=1}^{n} (x_i + y_i) &= (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n) \\ &= (x_1 + x_2 + \ldots + x_n) + (y_1 + y_2 + \ldots + y_n) \\ &= \sum x_i + \sum y_i \end{aligned}$$

Bisher diente als Laufindex der Fallindex i, der sich aus der Durchnumerierung der Untersuchungseinheiten i = 1, 2, ..., n ergab. Die Daten werden oft nicht nur nach diesem, sondern zusätzlich nach anderen Kriterien geordnet. Dann muß man zwei oder mehr Laufindices verwenden. So haben wir z. B. in Kap. 4 die Reichstagsabgeordneten nach ihrer Fraktionszugehörigkeit j = 1 (SPD), j = 2 (Linksliberale), ..., j = m = 5 (Konservative) geordnet und **innerhalb** jeder Fraktion den Fallindex i = 1,2,...,$n_j$ vergeben. Wenn wir z. B. das Lebensalter (X) aller Abgeordneten über alle Fraktionen hinweg addieren wollen (die Fälle aber nur innerhalb der Gruppen durchnumeriert sind), können wir diese Operation wie folgt ausdrücken:

**Regel 4:**

$$\sum_{i=1}^{n_j} \sum_{j=1}^{m} = \sum_{j=1}^{m} (x_{1j} + x_{2j} + \ldots + x_{n_j j})$$

$$= (x_{11} + x_{21} + \ldots + x_{n_1 1}) +$$
$$(x_{12} + x_{22} + \ldots + x_{n_2 2}) +$$
$$(x_{1m} + x_{2m} + \ldots + x_{n_m m}) +$$

Man hält beim Addieren zunächst einen Index fest (hier den Gruppenindex j = 1) und »durchläuft« den anderen Index (hier den Personenindex i = 1,2,...,$n_1$). Dann setzt man den zunächst festgehaltenen Index um eine Stufe nach oben (hier j = 2), hält ihn dort wieder fest und durchläuft wiederum den anderen Index (i = 1, 2,...,$n_2$). So verfährt man, bis der immer wieder »angehaltene« Index seine letzte Stufe (j = m, hier m = 5) erreicht hat.

**Regel 5:**

$$\sum_{i=1}^{n} \sum_{j=1}^{m} X_i \, Y_j = \sum_{i=1}^{n} X_i \sum_{j=1}^{m} Y_j$$

## Anhang:

### Themenübersicht zu Teil II

6. KAPITEL: Wahrscheinlichkeitstheoretische Grundlagen der induktiven Statistik

- 6.1 Zufallsexperiment und Zufallsvariable
- 6.2 Zum Wahrscheinlichkeitsbegriff
- 6.3 Das Rechnen mit Wahrscheinlichkeiten
- 6.4 Exkurs: Permutationen und Kombinationen (*)
- 6.5 Wahrscheinlichkeitsverteilungen und ihre Kennwerte

7. KAPITEL: Stichprobenfunktionen und ihre Verteilungen

- 7.1 Zum Konzept der Stichprobenfunktion und der Stichprobenverteilung
- 7.2 Binomial- und Multinomialverteilung
- 7.3 Die Normalverteilung
- 7.4 Von der Normalverteilung abgeleitete Verteilungsmodelle
  - 7.4.1 Die Chi-Quadrat-Verteilung
  - 7.4.2 Die t-Verteilung
  - 7.4.3 Die F-Verteilung

8. KAPITEL: Schätzen und Testen

- 8.1 Erstes Beispiel: Intervallschätzung des arithm. Mittels
- 8.2 Zweites Beispiel: Test auf »Signifikanz« einer Mittelwertdifferenz
- 8.3 Wünschenswerte Eigenschaften von Schätzfunktionen
- 8.4 Schätzen von Konfidenzintervallen
- 8.5 Zur Logik des Testens von Hypothesen
  - 8.5.1 Formulierung von Forschungs- und Nullhypothese
  - 8.5.2 Fehlertypen und Signifikanzniveau
  - 8.5.3 »Stärke« eines Tests
  - 8.5.4 Einseitige und zweiseitige Hypothesentests
- 8.6 Parametrische und nichtparametrische Testverfahren
- 8.7 Weitere Anwendungsbeispiele zu einzelnen Testverfahren
  - 8.7.1 Anteilsdifferenzen
  - 8.7.2 Der Chi-Quadrat-Unabhängigkeitstest
  - 8.7.3 Test auf Signifikanz des Pearsonschen Korrelationskoeffizienten r (*)
  - 8.7.4 Signifikanz von PRE-Maßzahlen für ordinale Variablen

## Literaturverzeichnis


**BERK**, Richard A., An introduction to sample selection bias in sociological data, in: American Sociological Review, Vol. 48 *(1983)*, S. 386 - 398.

**BERK**, Richard A./Subhash C. **RAY**, Selection biases in sociological data, in: Social Science Research, Vol. 11 *(1982)*, S. 352 - 398.

**BEST**, Heinrich, Struktur und Handeln parlamentarischer Führungsgruppen in Deutschland und Frankreich 1848/49 (Habilitationsschrift Universität zu Köln), Köln *1986* (Buchpublikation Düsseldorf 1989 unter dem Obertitel »Die Männer von Bildung und Besitz«).

**BLALOCK**, Hubert, Social statistics, New York usw. *1960*.

**BENNINGHAUS**, Hans, Deskriptive Statistik, 2. Auflage, Stuttgart *1976*.

**BORTZ**, Jürgen, Lehrbuch der Statistik für Sozialwissenschaftler, Berlin/Heidelberg/New York, *1979*.

**DAVIS**, James A., Elementary survey analysis, Englewood Cliffs, N. J., *1971*.

**FALTER**, Jürgen W./Kurt **ULBRICHT**, Zur Kausalanalyse qualitativer Daten. Grundlagen, Theorie und Anwendungen in Wahlforschung und Hochschuldidaktik, Frankfurt a. M. *1982*.

**FLOUD**, Roderick, Einführung in quantitative Methoden für Historiker, Stuttgart *1980* (deutsche Übersetzung der 2. Aufl. von: An introduction to quantitative methods for historians, London 1979).

**GIESEN**, Bern/Michael **SCHMID**, Basale Soziologie: Wissenschaftstheorie, München *1976*.

**GUILFORD**, J. P., Psychometric methods, New York, Toronto, London *1954*.

**HARTUNG**, Joachim/Bärbel **ELPELT**/Karl-Heinz **KLÖSENER**, Statistik. Lehr- und Handbuch der angewandten Statistik, 5. Auflage, München und Wien *1986*.

**HELLEVIK**, Ottar, Introduction to causal analysis. Exploring survey data by crosstabulation, London, Boston, Sydney *1984*.

**HILDEBRAND**, David K./James D. **LAING**/Howard **ROSENTHAL**, Analysis of ordinal data, Beverly Hills, London, New Delhi *1977*.

**HILDEBRAND**, David K./James D. **LAING**/Howard **ROSENTHAL**, Prediction analysis of cross classifications, New York usw. *1977 a*.

**HIRSCHI**, Travis/Hanan C. **SELVIN**, Principles of survey analysis, New York und London *1973*.

**JARAUSCH**, Konrad. / Gerhard **ARMINGER** / Manfred **THALLER**, Quantitative Methoden in der Geschichtswissenschaft. Eine Einführung in die Forschung, Datenverarbeitung und Statistik, Darmstadt *1985*.

**KENNY**, David A., Correlation and causality, New York usw. *1979*.

**KIRSCHNER**, Hans-Peter, ALLBUS 1980: Stichprobenplan und Gewichtung, in: K. U. **MAYER**/P. **SCHMIDT** (Hg.), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, Frankfurt a. M., New York *1984*, S. 114 - 182.

**KRIZ**, Jürgen, Statistik in den Sozialwissenschaften, Reinbek b. Hamburg *1973*.

**KÜHNEL**, Steffen M./Michael **TERWEY**, Einflüsse sozialer Konfliktlinien auf das Wahlverhalten im gegenwärtigen Vierparteiensystem der Bundesrepublik (Ms.), Köln *1988*.

**LIEBETRAU**, Albert M., Measures of association, Beverly Hills, London, New Delhi *1983*.

**REYNOLDS**, H. T., Analysis of nominal data, Beverly Hills, London, New Delhi *1977*.

**REYNOLDS**, H. T., The analysis of cross-classifications, New York und London *1977 a*.

**ROSENBERG**, The logic of survey analysis, New York und London *1968*

**SCHUBÖ**, Werner/Hans-Martin **UEHLINGER**, SPSS$^x$. Handbuch der Programmversion 2.2, Stuttgart und New York 1986.

**SCHLITTGEN**, Rainer, Einführung in die Statistik. Analyse und Modellierung von Daten, München und Wien *1987*.

**SCHMIERER**, Christian, Tabellenanalyse, in: Kurt **HOLM** (Hg.), Die Befragung 2, München 1975.

**SCHNELL**, Rainer/Paul B. **HILL**/Elke **ESSER**, Methoden der empirischen Sozialforschung, München und Wien *1988*.

**SCHRÖDER**, Wilhelm H., Forschungsstrategien in der Historischen Sozialforschung. Skript für den Einführungskurs des ZHSF-Herbstseminars 1987 (Ms.), Köln 1987. (Kurzfassung in, Historical Social Research/Historische Sozialforschung Suppl. No. 1 (1988)).

**SOMERS**, R. H., An approach to multivariate analysis of ordinal data, in: American Sociological Review, Vol. 33 *(1968)*, S. 171 - 177.

**SPSS INC.**, SPSS^X User's Guide. 3rd edition, Chicago *1988*.

**STATISTISCHES BUNDESAMT (Hg.)**, Datenreport 1985. Schriftenreihe der Bundeszentrale für politische Bildung, Bonn *1985*.

**THOME**, Helmut, Wertorientierungen und Parteipräferenzen in der Berliner Wählerschaft. Ein Forschungsbericht. Presse- und Informationsstelle der Freien Universität Berlin, Berlin *1985*.

**WILSON**, Thomas P., A proportional-reduction-in-error interpretation for Kendall's tau-b., in: Social Forces, Vol 47 *(1969)*, S. 340 - 342.

**WILSON**, Thomas P., A critique of ordinal variables, in: Social Forces, Vol. 49 *(1971)*, S. 432 - 444.

## Register